Title

De novo protein design—From new structures to programmable functions

Journal

Cell, 187(3)

ISSN

0092-8674

Author

Kortemme, Tanja

Publication Date

2024-02-01

DOI

10.1016/j.cell.2023.12.028

HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2025 February 01.

Published in final edited form as:

Cell. 2024 February 01; 187(3): 526–544. doi:10.1016/j.cell.2023.12.028.

De novo protein design – from new structures to programmable functions

Tanja Kortemme[1,2,3]

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco; San Francisco, CA 94158, USA.

[2]Quantitative Biosciences Institute, University of California, San Francisco; San Francisco, CA 94158, USA.

[3]Chan Zuckerberg Biohub; San Francisco, CA 94158, USA.

Summary

Methods from artificial intelligence (AI) trained on large datasets of sequences and structures can now "write" proteins with new shapes and molecular functions *de novo*, without starting from proteins found in nature. In this perspective, I will discuss the state of the field of *de novo* protein design at the juncture of physics-based modeling approaches and AI. New protein folds and higher-order assemblies can be designed with considerable experimental success rates, and difficult problems requiring tunable control over protein conformations and precise shape complementarity for molecular recognition are coming into reach. Emerging approaches incorporate engineering principles – tunability, controllability, modularity – into the design process from the beginning. Exciting frontiers lie in deconstructing cellular functions with *de novo* proteins and, conversely, constructing synthetic cellular signaling from the ground up. As methods improve, many more challenges are unsolved.

eTOC/in brief

Advances in artificial intelligence are revolutionizing protein engineering and design. This Perspective discusses the concepts and approaches of *de novo* protein design, emerging challenges in designing structure and function, and the frontiers that lie ahead in deconstructing cellular processes with *de novo* proteins.

Introduction

Proteins can accelerate the speed of chemical reactions by many orders of magnitude, convert the energy of light into chemical energy, and regulate the myriads of processes

*Lead contact and corresponding author: tanjakortemme@gmail.com.

Declaration of Interest

The author declares no competing interests.

within cells and organisms with the level of accuracy and precision required to sustain life. Because of these powerful functions, natural proteins have long been an attractive target for molecular engineering. The goals of protein engineering range from understanding the mechanisms of molecular and cellular functions to harnessing proteins for practical applications in catalysis, biotechnology, and as precision tools in discovery science and medicine.

The field of protein design is now fundamentally – and practically - rethinking this approach. Rather than reengineering existing proteins, it is becoming possible to build proteins with intricate architectures and functions – as powerful as those in nature, but new and user-programmable – from the ground up. This is the concept of *de novo* design[1], designing proteins from engineering principles or blueprints, without relying on existing starting points found in nature.

One can of course ask, why would one build everything new, if one can borrow, reuse, and reprogram from nature, or even arrive at functions new to nature despite starting from existing proteins[2]? Indeed, the approach of evolving or recombining existing protein components for new functions has been incredibly successful[2,3], and *de novo* design has long lagged behind because of its apparent limitations. Designed proteins, if less active than their natural counterparts, have required extensive screening campaigns to improve activity, and many desired functions seemed out of reach[4]. But if we could design functional proteins completely *de novo*, from the ground up, without the idiosyncratic features of evolved proteins, there may be several distinct advantages (Fig. 1A). The most obvious one is to enable functions not yet seen in nature (and for which there are no obvious existing starting points for directed evolution). The second advantage is that *de novo* design could allow us to create proteins that integrate engineering principles – tunability, controllability, modularity – into the design process from the beginning. We could engineer *de novo* proteins *a priori* to be (i) tunable, such that it is easy to generate versions with precisely altered biochemical parameters, (ii) controllable, such that protein function is responsive to internal and external stimuli, and (iii) modular, such that we can integrate different functions easily into composite molecular machines and assemblies.

Artificial intelligence (AI) promises a considerable leap in enabling this vision for *de novo* design. Recent advances in the accuracy of protein structure prediction through deep learning[5–7] have profound influence on the inverse problem, protein design, and are changing how *de novo* design is conceptualized. Classical approaches to protein design first define a protein backbone structure at the atomic level and then find a sequence that is consistent with that structure[8]. Designing “function” adds a definition of the structure of an active site (typically the relative atomic positioning of key catalytic or binding residues) that is built into a designed protein “scaffold”. Much of the difficulty of designing function lies in the fact that the designed protein needs to adopt the desired functional site structure with extraordinary precision. Even deviations of less than 1 Angstrom in atomic positions can cause the design to fail (if we, for example, think of the precise geometric requirements of hydrogen bonds). Consequently, much of the method development – and the challenge – of *de novo* design focuses on generating proteins that precisely adopt the desired geometry

(specific conformational dynamics and their timescales are other key challenges that I will discuss further below).

In contrast, generative approaches from deep learning offer the possibility, in principle, of designing structure, sequence, and function at the same time. The key conceptual leap seems clear, as structure, sequence, and function are intimately linked. A series of engineering problems of increasing difficulty illustrate the progression of design approaches that are currently being explored (Fig. 1B): (1) If we had a blueprint of the overall architecture of a protein (say, a barrel), could we experimentally realize instances of that architecture that are geometrically diverse (say, barrels of different sizes)? (2) If we had a blueprint of the positions of the most important atoms of a functional site in a protein, could we build a protein around this functional site, without needing to specify the protein fold or architecture that may be optimal for that function? (3) If we just had a function we wanted to design, could we ask a deep learning model to produce both a functional site and a protein sequence and structure model that harbors this site at the same time? (4) Or could we even simply ask the computer to design a protein that functions as desired? The answer to the first two questions is already yes in principle, approaches for the third are in development, and other applications – and more – are coming within reach.

The excitement about these advances in deep learning applied to *de novo* design does not mean that all problems are solved. Much the opposite, the rapid succession of new methods and their emerging successes in applications shift the focus from simpler design goals to many, often unsolved, larger problems; key and long-standing challenges of accuracy and precision, consideration of protein dynamics and conformational landscapes, and the scale of design problems are increasingly important. I will organize this review around these key challenges, advances, current state, and future opportunities. I will begin with concepts and approaches of *de novo* protein design, followed by chapters on (i) frontiers in design of new protein structures, (ii) new molecular functions, (iii) *de novo* proteins interfacing with cellular functions, and (iv) an outlook discussing long-standing and new problems. I will highlight developments in *de novo* design primarily in the last 5 years; there are many excellent reviews of earlier milestones (see Table 1 for a non-exclusive list of topic-focused reviews).

Concepts and approaches of *de novo* protein design

For several decades, approaches to computational *de novo* protein design used physics-based approaches and atomistic representations, grounded in structural biology principles and rules derived from naturally occurring protein structures. Now, advances in artificial intelligence are leading to rapid changes in methods. Still, many key concepts of *de novo* design and important challenges apply to both physics- and AI-based strategies.

Computational protein design as an optimization problem—Computational protein design is most fundamentally formulated as an optimization problem (Fig. 2A): Given a desired structure (and function), design methods seek to predict an optimal sequence that stably adopts that structure (and has that function). *De novo* design, which I focus on

here, does not start from naturally occurring, evolved proteins but aims to expand the space of protein structures, sequences, and functions beyond those seen in nature.

A key challenge is that the space of potential new sequences and structures is vast, sparsely populated with folded and functional proteins, and poorly mapped. For example, for a small protein of 100 residues there are $20^{100} = \sim 10^{130}$ sequence possibilities when considering the 20 naturally occurring amino acid types. Since the number of possibilities is larger than the estimated number of atoms in the universe ($\sim 10^{80}$), trying (termed sampling) all these sequences and their possible structures is impossible. Instead, efficient search algorithms are needed to navigate the enormous space of possibilities. At the same time, there are in principle vast numbers of *de novo* proteins with new sequences, structures, and functions that could be found.

Because functional proteins are rare among all possibilities, we also need rapid methods to distinguish between successful and unsuccessful sequences using computed “scores”. Most design methods have used either empirical or physics-based scoring or energy functions[9] that aim to estimate protein stability typically by considering atomic packing interactions, hydrogen bonding and electrostatic interactions, and solvation terms. The key challenge is to balance accuracy with speed, and this compromise necessitates approximations. Several sophisticated and well-tested atomistic simulation methods exist that use molecular dynamics with physics-based energy functions or even quantum mechanical calculations. However, each design candidate needs to be evaluated much faster than typically possible with these methods or else the approach is unlikely to find any viable solutions, even computationally. Unfortunately, a step-wise approach, first using approximate scoring functions followed by more accurate refinement, has proven difficult because fast, highly approximate scoring function tend to poorly correlate with the true free energy of proteins. In contrast, statistical approaches that learn from evolutionary sequence patterns[10] and more recent machine learning approaches (discussed below) that take as input even larger amount of data from sequence repositories instead of physics-based scores are revolutionizing the task of finding experimentally viable sequences.

Still the most fundamental and generally unsolved problem is the design of function. As computational design is an optimization problem, we need a quantifiable definition of “function” to optimize towards. Herein lie several challenges. Most fundamentally, such as for an enzyme, we may not have a sufficiently precise description of the requirements for function – such as defined conformational dynamics or electrostatics in an active site – even if we could design these properties accurately (see a recent perspective on challenges in enzyme design[11]). There are often multiple requirements for function – such as protein stability, the ability to adopt several conformations in a catalytic cycle, their rates of interconversion, specific recognition of desired interaction partners and avoidance of others, and more. Moreover, functional requirements can involve trade-offs, such as activity at the cost of stability, and computational approaches for multi-objective optimization are needed to balance these competing objectives. Finally, our ability to engineer many of these requirements with sufficient accuracy and precision is still limited, a challenge that I will come back to in the chapter on *de novo* design of molecular functions further below. Dependent on the design goal and the availability of a suitable starting point (a naturally

occurring protein) with an activity related to the target function, directed evolution may be the method of choice because the complex optimization criteria are implicitly encoded in an experimental screen for function in the desired context; even novel functions can be reached[2]. On the other hand, the mechanism by which the resulting functional proteins operate may not always be clear, and these proteins could therefore ultimately be less tunable and engineerable if the effects of mutational changes cannot be predicted.

Sequence optimization with atomistic modeling: fixed and flexible backbone design—To make protein design tractable given the challenges of sampling and scoring described above, most design approaches make a key conceptual simplification[12]: They divide the design problem into two steps: the first step generates a protein structure backbone (without a defined sequence), and the second step optimizes a sequence given that backbone (Fig. 2A). The second problem, termed fixed backbone design, was tackled first.

A milestone in fixed backbone design was reached in 1997 with the first complete computational redesign of a backbone structure existing in nature, a 28-residue zinc finger protein[8]. The design used discrete sampling of amino acid side chains with different conformations and residue types, a physics-based scoring function, and a deterministic optimization algorithm that found the global minimum energy sequence. A next milestone was the first computational design of a protein fold not found in nature, Top7. The design process using the modeling program Rosetta to first generate a new protein backbone (I will explain how in a section on structure generation using atomistic modeling further below), followed by iterative cycles of (i) sequence design given a fixed backbone and (ii) backbone minimization given a fixed sequence[13]. The Top7 example illustrates a key concept: Protein backbones are not fixed but they change, albeit often only slightly, when we make sequence changes in design or when proteins perform their functions. Many approaches have been developed to take this backbone flexibility into account in the design process, either by (i) backbone minimization interleaved with fixed backbone design as in the Top7 example[13], by (ii) sampling small backbone adjustments during design[14,15], or (iii) by pre-generating backbone ensembles onto which sequences are designed and scored[16,17].

Sequence optimization with AI: learning the language of proteins—Increasingly deep learning methods are applied to protein sequence design. AI-based protein structure prediction methods have learned from the vast amount of information in the database of protein structures (PDB) and sequence information for those proteins. Applying similar concepts, protein sequence design methods can learn from the vast amount of information in sequence databases, including those for which there is no structural information.

There are now many different machine learning models that have been developed for protein sequence design and structure generation (for recent reviews see[18,19]). Typically, AI methods for sequence design are evaluated by the extent to which the sequences predicted by the model resemble known sequences. A common metric is native sequence recovery, the fraction of predicted amino acid types at each position that are identical to those found in a native (naturally occurring) reference sequence. I will primarily focus the discussion here on AI models that have been experimentally validated. Experimental validation is

essential to determine the true success of design methods because only one (or a few) incorrectly predicted amino acids in the core of a designed protein will result in catastrophic experimental failure but only a small decrease in native sequence recovery.

One class of machine learning models that has been successfully applied to protein design are large language models (examples include ProtGPT2[20], ESM-2[7] and ProGen[21]). These models are trained on predicting missing amino acid "letters" in a protein sequence (analogous to language models trained on predicting missing words in a sentence). Once trained, protein language models can generate new protein sequences (just as ChatGPT is trained on text and can generate new text). ESM-2[7] is a language model trained solely on sequences (not structures) that has been applied to designing new proteins that are stable and monomeric when experimentally tested[22]. Notably, these proteins are predicted to have diverse structures including ones dissimilar to naturally occurring proteins (albeit there are no experimentally determined structures of these designs to date). These results indicate that the model may have learned an underlying grammar of proteins that generalizes beyond the training examples. ProGen[21] was similarly trained solely on protein sequences, but in this case from >19,000 protein families including labels of functional properties. For experimental evaluation, ProGen was fine-tuned on enzyme families (or a curated enzyme dataset) to generate designed variants with catalytic parameters similar to the natural proteins, including several with low (down to ~31%) sequence identity to any protein in the training set. Like ESM-2, ProGen does not require a protein structure for design but does require large datasets of sequences for a given protein family. Analogously, a previous machine learning model, UniRep[23], was shown to predict functional properties of proteins to enable variant engineering when fine-tuned on appropriate datasets. A different study showed that language models can be adapted for design of diverse functional sequences without the need for sequence alignments[24]. This method successfully generated diverse, well-expressing nanobodies for which alignments are difficult because of high diversity in loop lengths and sequences. Language models were also successfully applied to model-guided affinity maturation of antibodies[25].

Other models for sequence design take both sequence and three-dimensional structure as input. Given a fixed protein backbone, these models predict amino acid identities using the local structural environment as context[26] (sometimes represented as a graph[27,28]). ProteinMPNN[28] builds on a prior model for graph-based protein design[27] and has been extensively validated experimentally on designing proteins with existing and novel folds and large symmetrical protein assemblies. In addition, the model has been fine-tuned to predict the effect of single amino acid point mutations (ThermoMPNN[29]) using large datasets of stability measurements[30]. Frame2seq[31] is a recent model that, in contrast to ProteinMPNN, predicts sequences in a single pass with increased or comparable accuracy but improved speed and a score that reflects prediction accuracy. One important question is to what extent deep learning models generalize, i.e. make predictions outside of the datasets they are trained on. Here, experimental validation suggests that Frame2seq may be able to design stable proteins with undetectable similarity to the starting protein, allowing exploration of novel sequence space. Overall, the high success rates of AI-based sequence design methods in experimental validation (often >10%, in favorable cases >50%) vastly increases the number and types of applications addressable with computational design.

Structure generation using atomistic modeling: design of all major fold classes and symmetrical assemblies—Experimentally validated, state-of-the-art models for *de novo* protein sequence design, such as ProteinMPNN[28] and Frame2seq[31], require a protein backbone as input. This requirement poses two problems. First, one needs to have a method to generate new protein backbone conformations (Fig. 2). Second, one needs to assess whether these backbones are "designable", meaning that there exists at least one sequence that stably folds into that structure.

The most obvious way to fulfill the designability criterion is to start with a protein backbone conformation existing in nature and repurposing it for a new function. Indeed, this approach has been successful in many cases. For example, computational design approaches have been developed to redesign enzymes for altered substrate specificity[32] and protein-protein interfaces for orthogonal signaling[33]. However, seemingly straightforward changes in specificity can be surprisingly difficult to design with computational approaches. A primary reason is the limitation given by the starting backbone conformation. For example, simply replacing a hydrophobic with a polar side chain to interact with a more polar substrate may not place the polar functional group in precisely the correct geometry for optimal hydrogen bonding with the new substrate, and even small deviations in geometry can have detrimental effects on function. For these engineering problems with a close starting point, directed evolution strategies are more suitable.

For generating protein backbone conformations *de novo*, the problem of designability can be solved in very elegant ways for all-helical structures. Here, breakthroughs were made when applying a set of parametric equations describing the geometry and relative orientation of interacting helices (Crick's parameterization), which make it straightforward to generate large sets of designable helical coiled-coils. Extensive design and experimental validation studies led to a systematic description of a "periodic table" of coiled-coil architectures[34]. Crick's parameterization can be extended to arbitrary helical bundle architectures[35] that, when designed and tested in the laboratory, can be extremely thermostable[36]. Moreover, helical architectures can be spliced together[37]: The regular geometry of helices allows the alignment of helices in different proteins, leading to a facile method to generate a range of structurally distinct proteins[37] and larger helical architectures through fusion of overlapping helical regions. Helical repeat proteins with different curvature[38] then allow design of large assemblies with an impressive systematic variation in geometries[39]. The diversity of designable all-helical structures still underlies many of the successful applications of *de novo* designs[4,40,41]. However, while the problem of designing alpha helical proteins is largely solved due to our understanding of the design rules, more complex functions may require more structurally diverse structures with deviations from canonical helical geometries.

Much progress has also been made with the *de novo* design of protein folds containing a mixture of alpha helices and beta strands. A typical design process follows a four-step approach: The first step defines a "blueprint" of the desired protein fold topology, defined as the identity and connectivity of alpha-helical and beta-strand secondary structure elements (Fig. 2B). Blueprints allow for the definition of new fold topologies not found in nature[13]. The second step is to assemble a protein backbone from peptide fragments

(helices and strands) according to the blueprint and connected by short loops (Fig. 2B). Peptide and loop fragments are typically taken from overrepresented fragments in the PDB, thus ensuring designability at least at the level of local (one-dimensional) sequence-structure compatibility[13]. Designability at the fold level can be assessed by rules found in existing protein topologies, such as organization of secondary structure elements into tertiary motifs[42]. An impressive example was the *de novo* design of symmetrical TIM-barrel proteins[43], a long-standing challenge in design that required specific side chain-backbone hydrogen bonds for defining the strand register between the barrel repeat units to succeed. The third step involves sequence design, often iterated with backbone minimization, as described for Top7[13] above. This step generates sequences predicted to be optimal for the desired input structure. A final step assesses designed sequences in *silico* by predicting their structures and comparing the prediction to the intended backbone. Designs passing this test are experimentally validated. These approaches led to the design of diverse alpha-beta protein folds[38], and were generalized in methods such as TopoBuilder[44].

The design of structures with exclusively beta-sheet secondary structures (all-beta proteins) poses distinct challenges. For example, all-beta proteins show a tendency to aggregate. Moreover, attempts to derive parametric design methods, such as for helical bundles, have not been successful. Instead, breakthroughs were made through the realization that beta-barrel structures in nature have defined defects that allow relief of strain that would be present in idealized barrels[45]. This principle allowed the design of a range of beta-barrel geometries and a functional fluorescence-activating beta-barrel[45]. Other design efforts have generated beta-sandwich folds[46,47].

In addition to generating new tertiary structures with different folds, computational design has also been applied to generate quaternary structures. Particularly exciting are the designs of a large variety of symmetrical assemblies with impressive sizes, with important applications as delivery vehicles, reaction compartments, or nanoparticles for vaccines[48]. Designing these assemblies typically involves docking of the component (natural or *de novo*) monomers in the desired symmetry, and redesign of the resulting interfaces. The design of these architectures is aided by symmetry: any designed interface interaction (if net favorable) will be repeated many times in the assembly, adding up to overall stabilization.

All the structure generation methods discussed above require a desired target structure or blueprint that needs to be prespecified at the start of the design process. The AI-based structure generation methods described in the next section do not have that requirement, opening up new avenues for the formulation of design problems.

Structure generation using AI models: natural and novel folds—The breakthroughs in AI-based methods for protein structure prediction, such as Alphafold2[5], trRosetta[49] and RoseTTAFold[6], have inspired numerous recent advances to invert these models for design: Instead of predicting a structure given a sequence, the task is to *generate* a structure from scratch and then predict its sequence (methods that generate sequences and structures at the same time are less explored at present). One of the key differences to the parametric or blueprint-based structure generation methods in the previous section is that

AI-based methods do not necessarily require definition of the desired protein structure or fold class *a priori*.

Among the first AI-based approaches that were experimentally validated by *de novo* design is protein "hallucination"[50] that inverts the trRosetta structure prediction model for structure generation. Here, sequences are optimized to adopt predicted tertiary structure contact maps that resemble those of natural proteins but are different from those of random sequences. While hallucination generates both backbones and corresponding sequences, many hallucinated designs were not successful when tested experimentally. Considerably higher design success rates were reached when the hallucinated backbones were redesigned with ProteinMPNN in a second step[28]. The necessity of this second step may reflect the insensitivity of current protein structure prediction methods to amino acid point mutations that can be catastrophic in protein design. Hallucination has been used to generate proteins and symmetrical assemblies with experimentally validated structures[50,51].

More recent AI-based protein design strategies use diffusion models[52–54] borrowed from image generation. Diffusion models start with images that are successively "noised", followed by training a network on the noised samples to recover the original images. In the case of proteins, diffusion models start with protein structures and add successive noise to the protein coordinates, followed by training to recover the original structures. Using these models for design, one starts from random noise, and the denoising process generates samples of protein structures with properties of those resembling typical proteins (Fig. 2D). One such model, RFdiffusion[53], has been used to generate experimentally validated protein monomers, symmetrical assemblies and protein binders, and appears to outperform hallucination-based approaches. Another diffusion model, Chroma[54], has been used to generate experimentally validated protein monomers. A particularly exciting property of diffusion models is that they can be conditioned in various ways, such as generating particular fold topologies (Fig. 1B – 1) or preserving specified functional sites (Fig. 1B – 2), applications that will be discussed in the chapter on *de novo* design of molecular functions below.

Frontiers in design of new protein structures

As outlined above, proof-of-principle studies in *de novo* protein design have built diverse representatives of the major secondary structure architectures of proteins (all-alpha, mixed alpha-beta, and all-beta) as well as impressive higher-order symmetrical assemblies of them. Moreover, new protein structures can now be generated with considerable experimental success rates[4] (often > 10%), with further increases through the development of recent AI models for both structure generation and sequence design. In this chapter I will focus on frontiers in design of protein structures. I will describe approaches to explore novel fold space, test mechanistic principles through reengineering them, and engineer user-defined shapes tunable for new protein functions. Together, these design strategies begin to build a framework for the *de novo* design of complex architectures and molecular machines.

Principles through bottom-up construction—While naturally occurring proteins occupy a limited number of protein topologies or folds, the early design success of Top7[13]

demonstrated that a stable new topology not seen in nature could be generated through computational methods. Generalizing this idea, a systematic exploration of alpha-helical coiled-coils led to the design of novel architectures and development of principles to exploit these architectures for diverse functions[55]. Recent advances in AI-based computational protein design now allow in principle to map protein fold space systematically. New backbone generation methods such as GENESIS[56] are being developed to do so, and could be used to generate novel folds likened to cosmological “dark matter”. Ultimately, more systematic maps of protein fold space could (beyond generating starting materials for engineering) allow for better quantification of designability principles and thereby advance the speed and accuracy of design.

In naturally occurring proteins, functional mechanisms are often coupled in complex ways, reflecting aspects of the history and context in which functions evolved. In contrast, building new functions from the ground up might allow the dissection of principles that are difficult to entangle in evolved systems, such as principles of conformational switching, allosteric control, or mechanical stability. Designing these complex functions *de novo* is a difficult problem currently but could be reachable in the future.

Finally, a key frontier is the ability to dissect quantitative determinants not only of molecular, but also of cellular, tissue, and organismal functions. Here, *de novo* designed proteins could be engineered to have precise and systematic variation of molecular properties that in turn tune higher-order biological processes (I will come back to this aspect in a chapter on *de novo* proteins for cellular functions below).

Precise control over protein geometries: Synthetic fold families for function
—Nature does not invent a new protein fold for every new protein function. Instead, existing protein folds are customized and optimized for new functions through changes in fine-grained geometries of functional sites and tuning of relevant protein dynamics. To design biological functions with biologically useful activity and required accuracy, computational design should therefore be able to exert precise control over fold shape as well as functional site geometry and dynamics. Considerable progress has been made with controlling overall course-grained variation of protein folds, as described above. In this section, I will highlight advances with developing methods that allow fine-grained control over the precise geometries of proteins to optimize details of atom-level interactions in functional sites. I define “geometry” as the variation or features including length and orientations of secondary structure elements within a given fold topology (the identity and connectivity of secondary structure elements).

The blueprint structure generation methods (Fig. 2B) described earlier typically generate idealized versions of the targeted fold topology, and although thousands of stable variants can be designed[57], they often are very similar to each other (1–2 Å root mean squared deviation, RMSD). Several approaches have been developed to instead systematically sample fine-grained geometrical features[58–60] such as pocket shapes[59]. Since a large fraction of evolutionary variation involves diversity in positioning of helical elements, the LUCS sampling method[58] enables generation of synthetic fold families with tunable geometries through systematic variation of position, orientation, and lengths of helices (Fig. 2C).

Several experimentally determined structures showed how the *de novo* designed proteins with identical fold topology can have large diversity in geometry, in each case in excellent atomistic agreement with the design model. The ability to in principle sample thousands of finely tunable geometries should allow progress with another frontier: the design of defined dynamics and conformational changes (discussed further below).

Complex shapes and blueprints for protein machines—The ability to generate larger protein structures through helical fusions and controllable oligomeric assemblies opens up new avenues to engineer more complex architectures with arbitrary shapes. These shapes could be, for example, the parts of molecular machines and motors (such as rotors and axels), which would need to break symmetry to undergo motion (rotation around the axel). A fascinating example of the design of diverse synthetic protein-based rotor and axel components and their assembly to prototype protein nanomachines[61] was recently described. There are many open challenges such as driving rotation through energy conversion using chemical fuels.

Further advances in AI-based methods might allow design of complex protein shapes for nanoscale machines and biological patterns by first drawing a component blueprint and assembly plan, followed by custom-optimization of the required protein shapes. In addition, the design could consider the engineering principle of modularity during the design process of these larger assemblies so that they can be built up from plug-and-play pieces.

De novo design of molecular functions

The progress made with the accurate *de novo* design of new protein folds and diversified shapes and geometries, with success rates approaching >10% or even >50% dependent on the design goal[4], contrasts with the ongoing challenge of designing new protein functions. Typically, computationally designed proteins provided a starting point with robustly measurable but low activity that would subsequently need to be optimized experimentally to achieve practically useful functions. With the advances of deep learning methods, this paradigm is beginning to change, at least for an initial range of functions. I will first highlight general principles of computational design of function, then outline how AI-based methods are changing the process, and finally describe state-of-the-art applications and frontiers.

Principles for designing function: motifs and scaffolds—Most generally, computational design of function (Fig. 3) involves two steps: The first step defines the requirements for function and the second step optimizes a protein structure and sequence that matches these requirements. With advances in deep learning applied to proteins, how these steps are carried out is changing rapidly, increasingly with notable success rates.

Most computational approaches to date define the requirements for function as precise and pre-organized active site geometries (Fig. 3A, B). More specifically, these geometries are often defined as the relative position and orientation of functional groups of amino acid residues in a protein active site – for example the positioning of an arginine guanidinium group in suitable hydrogen-bonding geometry and distance to a carboxylate on a protein or

small molecule binding partner. The key challenge then is to achieve this precise positioning for multiple interacting groups in a functional site stably designed into a protein scaffold.

Initial successful applications of this concept defined a few functional site geometries, also called motifs, either by rational design of active site interactions or by borrowing motifs from natural proteins, and then transplanted (matched) the motif into a different naturally occurring protein that was used as scaffold[62]. These approaches are principally limited in several ways: First, the precision with which any motif can be accommodated in a scaffold is intrinsically constrained by the available (natural protein) scaffold backbones. To optimize the motif precision, only small adjustments to the backbone were possible in earlier design processes. As a result, the desired geometry was never placed exactly in the desired geometry, often resulting in loss of function. Second, naturally occurring proteins are often only marginally stable. Placing a new functional site into them can therefore lead to unfolding. Third, more complex functional sites with more than 3 to 4 residues can frequently not be matched with reasonable precision to any natural scaffold.

The first and second problems can be addressed by using libraries of *de novo* designed proteins as scaffolds (Fig. 3B). Approaches where scaffolds can be finely tuned in their geometries, such as helical bundles through parameterization[63] or other folds through the structure diversification approaches[58–60] described above, are particularly successful. In addition, *de novo* designed proteins are often extremely stable, overcoming issues with placing functional sites into them.

The problem of not finding any suitable matches in a given library of pre-generated *de novo* scaffolds is more complex. To a certain extent, this problem can be overcome by increasing the numbers: generating tens of thousands of potential motifs through computational methods[64,65], and matching these into libraries of hundreds or thousands of scaffolds[65,66].

However, more general approaches that optimize (or even generate) the protein scaffold given a functional site definition, are necessary. Solutions to this “motif scaffolding problem” are in active development using various AI models for proteins. For example, given a motif geometry as input, both protein hallucination[67] and diffusion[53] can in principle generate a suitable scaffold around that motif for a range of scaffolding problems (Fig. 1B – 2; Fig. 3C). The key challenges here are in assessing (i) that the generated protein backbone is indeed designable and (ii) that the precarious details of non-covalent interactions are sufficiently accurate to stabilize the functional site in its desired geometry. Both criteria are currently assessed by predicting the structure of the generated design sequence using an orthogonal deep-learning method that was not used in backbone generation. While a useful computational consistency check, these methods can be insensitive to the effect of small details of interactions. Moreover, most of these methods currently do not explicitly model any non-protein ligands. Nevertheless, the reported success rates with these approaches in functional assays, as detailed for specific applications below, are impressive. Still, few functional designs generated by these methods to date have been validated by high-resolution experimental structures; further data are therefore needed to systematically assess the accuracy of designed functional site geometries.

Finally, AI-based methods should in principle also be applicable to the first step: definition of the requirements for function. An example is the MaSIF (molecular surface interaction fingerprinting) method that captures “chemical fingerprints” of suitable interaction interfaces on a protein target that can be computationally matched with complementary surfaces to generate *de novo* protein binders[66] (Fig. 3E). In a different approach, language models trained on protein families appear to encode requirements for function, because these models can be used successfully to generate designed variants with that function[21].

Molecular recognition: protein-protein interactions—The *de novo* design of protein binders recognizing target protein partners[68] has led to exciting applications such as selective cytokine mimics[69], and protein inhibitors of a histone methyl transferase[70] and the SARS-Cov2 spike protein[53]. The first approaches to computational binder design created new interfaces between existing proteins[71] and altered the specificity of existing interfaces[72]. A key development was “hotspot-directed” design[73], later generalized using a “rotamer interaction field” approach[45,65]: Here, disembodied amino acid side chains are docked against a target surface of interest to identify ideal interactions in a desired surface. In a second step, these docked side chains are incorporated in a scaffold protein to generate a binder (Fig. 3D), first using natural scaffolds and later *de novo* designed proteins. An impressive larger-scale design study assessed the success rates of this approach[65]. For a panel of 12 targets with different shapes, the computational approach could generate binders for all targets, with success rates for identifying binders in the micromolar range between <0.01 and 1% (using libraries of 15,000 to 100,000 design candidates per target). To achieve nanomolar to picomolar binding affinities, the binders (all hyperstable mini-proteins smaller than 65 amino acid residues) were subsequently optimized using mutational screening.

Recent AI-based methods to protein binder design constitute a step advance (Fig. 3E, F), leading to higher success rates without reliance on large libraries or extensive experimental optimization. For example, RFdiffusion was shown to generate binders in the micromolar range for 5 targets with a 19% estimated success rate, testing fewer than 100 designs per target[53]. For two targets, low nanomolar binders were identified with no further experimental optimization. Designs generated using the MaSIF method identified binders for 4 targets[66]. For one target, the method yielded a low nanomolar binder without experimental optimization. While the RFdiffusion study above used predefined interaction hotspots on the targets, AI methods such as MaSIF could also be applied to identify good interaction surfaces on targets for which there are no known interaction hotspots. Another promising approach applies iterative design and structure prediction cycles to refine initial designs in a process akin to *in silico* directed evolution[74]. The Sculptor[75] method uses deep learning to optimize the backbone conformation of a protein binder for a given target surface. This method addresses a long-standing challenge in computational design: to mimic the ability of antibodies to evolve high shape complementarity to many diverse targets by exploiting the conformational plasticity of loops. In addition, computational design methods such as Sculptor have the advantage over experimental antibody selection methods that the target surface can be specified *a priori*.

Despite significant advances in binder design, not all challenges are addressed. Key difficulties include the design of binders for target surfaces that are highly flexible or

very polar. Nevertheless, progress is being made with explicitly considering flexibility in molecular recognition[76] and biased design for polar contacts in the interface[66]. It will be interesting to analyze the growing number of successfully designed *de novo* binders for privileged interfaces or interaction modes. As yet, helical interaction surfaces on the designed binder are overrepresented (although not exclusively). Helices are more designable owing to their regular geometries, well-known design rules, and the intrinsic property that backbone donor and acceptor groups are internally satisfied; hence, the detrimental effect of unsatisfied buried hydrogen bonding donors and acceptors in helical interface is minimized.

Molecular recognition: protein-small molecule interactions—Small-molecule recognition is key to numerous protein functions including catalysis and signaling. Design of proteins binding to small molecules has remained a difficult problem, with few examples of engineering small molecule binding sites *de novo* into existing proteins[77,78], as well as *de novo* designed helical bundles[63] and a beta-barrel[45]. In particular, highly polar or flexible small molecules are more challenging targets due to the difficulty of optimizing the precise geometries of polar contacts or the entropic penalties incurred when binding ligands with many rotatable bonds. Overall, the achieved affinities are typically in the micromolar or high nanomolar range before experimental optimization. Nevertheless, these approaches offer exciting opportunities for design of small-molecule-induced assemblies to control extra- and intracellular signaling processes.

Several deep-learning approaches have been proposed to scaffold motifs for interactions with small molecules. To date, many studies report *in silico* benchmarks. Experimental success (although no experimentally determined structures) has been reported for scaffolding metal binding sites[53]. Very recently, an all-atom version of RFdiffusion, RFdiffusionAA,[79] has been applied to design proteins binding to the therapeutic digoxigenin, the enzymatic cofactor heme, and other targets. For digoxigenin, ~4,400 designs were experimentally screened to identify three designs that showed enrichment in a yeast display assay, with one design binding in the nanomolar range. While these are currently modest success rates, an exciting aspect of the method is that it could achieve high shape complementarity to small molecules by simply defining the target without having to pre-generate a binding motif. It will be interesting to compare the agreement between the AI-generated design models and experimentally determined structures for these emerging design methods.

Multi-objective optimization: Conformational changes and switches—The functions of evolved proteins are typically complex and composite, such as coupling binding to conformational changes, or posttranslational modifications to changes in activity. To ultimately match and surpass the advanced functions of natural proteins, *de novo* design approaches must be able to optimize over these different objectives. Such approaches are at early stages, with some notable advances.

One frontier area is to design tunable conformational switches by optimizing single sequences over multiple conformational states. Pioneering examples led to the design of a protein that switches between two different secondary structures[80] and proteins that have different designed conformations distinguished by alternative states of a tryptophan side

chain moving on the millisecond time scale[81]. Most recently, switches have been designed that upon peptide binding interconvert between two different structured states related by an overall hinge-motion of two helical subdomains[82]. This latter application was enabled by the ability of the AI-based sequence design model ProteinMPNN[28] to optimize sequences while simultaneously considering two different structures. In the case of hinge-proteins, the problem is simplified since most intramolecular interactions stay the same except certain intra-domain interactions altered by the hinge.

For some naturally occurring protein switches, AlphaFold2 can recapitulate alternative states among the different generated model predictions[83]. It is an open question to what extent AI-based structure prediction methods can predict and design multiple states *de novo*, without having been trained on natural examples of a given conformational switch.

Ideally, computational methods should be able to accurately predict the underlying distributions of conformations, and efforts to develop such methods are underway[84,85]. It will be exciting to see applications of these concepts to the *de novo* design of conformational switches and other advanced functions that require explicit consideration of conformational changes or allosteric effects[76]. The area of multi-objective designs of conformational switches is likely to see further advances in the design of more complex, composite protein functions *de novo*.

De novo proteins for cellular functions

Synthetic signaling systems that can control biological processes (chimeric antigen receptors are a prominent example) have many significant applications in fundamental biology, bioengineering, and medicine. The vast majority of such signaling systems built to date have used naturally occurring components (genetic elements and proteins) and recombined or reprogrammed them for new functions[3,86]. The increasing success of *de novo* protein design now allows, in principle, to build protein signaling systems entirely from the ground up. Unlike natural proteins that are evolved to function in specific contexts, *de novo* proteins could be engineered *a priori* with context-independent function that allow tunability and modular behavior (Fig. 1). In addition, new functions not yet seen in nature may become accessible. *De novo* proteins could be engineered to sense new signals, integrate signals and perform logic, and precisely regulate downstream biological behaviors (Fig. 4). For each of these functions, computational methods could generate elementary components with tunable properties (such as binding on- and off-kinetics, diverse assembly geometries, etc.), and these components could be linked together in a modular fashion to generate diverse signaling behaviors. In this section I will describe progress with computational engineering of proteins for cellular functions, from reprogramming existing proteins to designing components *de novo*. I will also highlight how engineering principles of modularity and tunability are being incorporated into the design process, and how designs are interfaced with cellular processes to dissect principles of regulation.

Design of sensors and actuators with diverse inputs and tunable outputs—The ability to sense and respond to molecular signals is a fundamental ability of all living systems, and engineering it *de novo* could advance many areas of science, technology,

and medicine. Examples include metabolic engineering, by monitoring intermediates in production of industrially valuable chemicals; cell signaling, by creating tools to dissect normal and disease processes with improved precision; and cancer treatment, by achieving tight regulation of advanced therapies such as CAR-T-cells. An exciting example of a computationally designed sensor that functions at the organism level to track the distribution of the plant signaling molecule auxin in plant roots in real time was recently described[87] (Fig. 4A).

A key challenge in designing new sensor/actuator systems is to develop generalizable ways to couple detection (sensing) of a signal to a cellular output response (actuation). Unless the signal is intracellular or readily traverses a cell membrane, engineered sensor/actuator systems must transmit the signal from the outside of the cell to the inside. No entirely *de novo* engineered transmembrane signaling system exists yet. Nevertheless, progress has been made with reengineering existing transmembrane signaling systems to modulate allosteric signal transduction[88] and quaternary structure changes[89] in GPCRs.

Ideally, an engineered system should be specific to the signal but modular in its output response, such that a given input signal can be linked to a variety of output responses that could be changed without having to reengineer the entire system. One architecture that fulfils these criteria is chemically-induced heterodimerization (CID). Here, two components of a sensor preferentially heterodimerize in the presence of a small molecule, which can be linked in a modular fashion to complementation of a functional output reporter. Many suitable split reporters exist that activate, for example, fluorescence, enzyme activity, or most generally expression of any gene or gene combination. CID systems can be entirely intracellular, but can also provide a coupling mechanism across the membrane when sensing triggers the preferential assembly of transmembrane proteins with domains on either side of the membrane. Several CID systems have been rationally engineered based on selecting binders to drug-bound proteins as starting points[90–92]. To date, one modular sense/response system has been built by *de novo* computational design of a small molecule recognition site[78] (Fig. 4A), albeit by engineering it into an existing protein-protein interface to create a CID system. The synthetic system showed dose-response behavior in cells to detect a metabolic intermediate produced via an engineered pathway. The output response could be exchanged in a modular fashion, and a crystal structure of the assembly showed good agreement with the computational design model.

Advances in computational design now pave the way to design CID systems with tunable binding behaviors entirely *de novo*. Moreover, the specific architecture of CID systems can determine different input/output behaviors (Fig. 4B). For example, CID systems can exhibit a “bandpass filter” response[93], where the signal is high only at intermediate signal concentrations but low otherwise. Other CID systems can show “molecular ratchet” responses that shift the response amplitude and sensitivity dependent on the concentrations of the CID components[94]. Modeling the quantitative response of different CID architectures creates exciting opportunities to realize different input / output behaviors with engineered systems. Looking into the future beyond CIDs, one could imagine creating *de novo* sensors and actuators for diverse inputs such as peptides, pH, light, ionic strength, temperature, and mechanical force.

Regulation and logic—Another key property of all living systems is the ability to integrate signals and make decisions. Cellular decision making takes place in complex signaling networks, where not all interactions and their functions are known. Synthetic signaling systems offer the advantage of simplifying feedback and regulatory mechanisms such that they can be finely tuned and robustly controlled.

A pioneering study engineered *de novo* helical bundle proteins such that they could be embedded into positive and negative feedback system controlling both natural signaling (the yeast mating pathway) and synthetic gene circuits[95]. The regulation mechanisms were based on a protein domain replacement strategy in the *de novo* designed LOCKR (latching orthogonal cage-key proteins) system[96] (Fig. 4C). Here, an output element located on a helix is buried inside a *de novo* helical bundle but can be displaced by a “key”, a helical input element, that competes with the locked helix. Feedback mechanisms could be engineered by designing a degronLOCKR, where the output element is a protein motif important in regulation of degradation (degron). The degron is exposed in the presence of the input signal (the key) and targets a fused cargo protein to the proteasome. The system was shown to be tunable, even in a combinatorial fashion, by modulating the key’s production (via an inducible promoter) or the key’s binding affinity to the degronLOCKR (by changing the length of the key).

A different study implemented colocalization-dependent regulation (Co-LOCKR) that perform AND, OR, and NOT Boolean logic operations in response to combinations of molecules present at the cell surface[97] (Fig. 4D). Other *de novo* proteins that have been used to implement Boolean logic in cells include sets of helical bundle heterodimers with engineered specificities mediated by hydrogen bonding networks linked to a split luciferase reporter or transcriptional regulators[98] and designed coiled-coil dimerization domains linked to split proteases[99].

Self-assembly and localization in cells—There has been long-standing interest in the signaling properties of cellular assemblies, ranging from higher-order oligomers to membrane-less compartments. Engineering such systems *de novo* could both contribute to deconstructing the function of natural systems as well as exploit specific characteristics such as signal amplification. Efforts to engineer *de novo* proteins that self-assemble in cells are beginning to emerge. For example, *de novo* helical proteins were designed to assemble into membraneless organelles whose assembly dynamics can be controlled. One assembly was shown to co-compartmentalize an enzyme pair to improve product formation[100]. In a second example, pairs of *de novo* designed symmetric protein homo-oligomers each comprising 2–120 individual protein components were shown to assemble in mammalian cells into protein networks whose mechanical properties could be tuned intracellularly[101]. A third study designed *de novo* single-pass alpha-helical transmembrane domains that assemble into defined dimers, trimers, and tetramers[102]. These and similar designs could be used to probe how defined changes in valences and geometries of protein signaling assemblies affect biological responses.

Other approaches are beginning to engineer cellular delivery by designing *de novo* binders to transmembrane receptors triggering endocytosis[103]. Another study developed a *de novo*

designed system with dual function for both delivery and subcellular localization[104]. Here, an arginine-rich peptide is designed to penetrate the cell and subsequently bind a complementary acidic partner that can be fused to various other proteins to control subcellular localization.

Interfacing with and deconstructing biological processes—Ultimately, to deconstruct and regulate complex biological processes, *de novo* engineered systems must have robust interfaces with complex biological machinery. One way to do so is to use *de novo* designed proteins as assembly parts for downstream biological processes. Here the *de novo* components could provide controllable inputs (such as the CID systems discussed above[78,92]), tunable assembly kinetics, or defined geometries. Design efforts with these goals are beginning to emerge and provide new tools to probe necessary and sufficient parts of natural signaling. For example, extracellular 2-dimensional arrays of *de novo* designed proteins have been used as assembly parts[105] linked to intracellular proteins of interest via transmembrane helices. Inducible extracellular assembly promoted intracellular clustering, which was then used to trigger polarity of protein targets in mammalian cells and dissect regulatory events sufficient for cytoskeleton polarization. In another example, *de novo* designed proteins were used to change valences and geometries of synthetic cell surface receptor ligands. Here, *de novo* designed cyclic homo-oligomers with up to eight subunits were linked to a *de novo* designed FGF receptor binding protein and applied to probe and manipulate FGF signaling[106]. Notably, defined oligomers are uniquely engineerable with designed proteins, in contrast to standard antibody reagents or natural binders.

Engineering principles—Increasingly, *de novo* design studies adopt strategies to engineer protein functions that can be readily expanded beyond a single example into families of *de novo* proteins that could be used as elementary components in engineering larger, compositive synthetic signaling systems. Consider for example, instead of building one sensor for a specific signal, building a family of sensors for that signal that have different input/output characteristics. Another example would be to engineer a set of signaling assemblies with the same architecture but controllable by different signals. A third example would be a set of *de novo* protein-protein interaction elements with different on- and off-kinetics or oligomeric assembly properties, which can be linked together in combinatorial and modular fashion. Ideally, all of these could be combined to construct signaling systems with desired “off-the-shelf” characteristics and not requiring extensive re-optimization in each specific context. Table 2 summarizes examples of emerging approaches to designing extendable systems to be tunable, controllable, and composable.

Challenges and next opportunities

Advances in AI are revolutionizing protein design, and new methods are emerging rapidly. Currently, successful experimental applications address relatively simple problems, such as design of idealized folds (still with an overrepresentation of all-helical proteins), symmetrical assemblies, and protein-protein interfaces – albeit most recently with examples of remarkable shape complementarity. The increasing success rates of these applications are bringing important, long-standing challenges, such as design of precise geometries of polar functional sites and dynamical proteins, into reach. Latest developments such as protein

diffusion models that model not just the protein backbone but all atoms including side chains and ligands[79] can be used to generate proteins *de novo* around small molecule ligands, albeit still requiring screening of relatively large numbers of designs. Further-reaching design goals such as molecular machines are coming into reach, and more complex composite functions can be deconstructed into designable components and implemented[61].

Deep learning and data—The step-change with AI-based protein structure prediction required vast datasets of protein structures and sequences. In principle, function is also encoded in these structures and sequences, and this encoding has been used by machine learning models to generate functional proteins[21,23–25]. However, precise requirements for specific target activities and dynamics are more difficult to extract for desired properties where we lack informative datasets. Herein lie both significant challenges and opportunities for advances reachable by deep learning. Integration with approaches for quantitative measurements of functional parameters at scale seems to be one promising avenue. There are exciting opportunities for new capabilities to generate robust and accurate large-scale datasets that validate designs and probe their stability[30,57], as well as recent high-throughput methods for rapid determination of rate constants and affinities[107].

Multiple objectives and energy landscapes—Advanced protein functions will most likely involve integration of properties, such as the cycles of molecular recognition, resulting conformational changes, and exposure of new recognition sites exhibited by naturally occurring protein switches (such as regulatory GTPases). More generally, diverse inputs modulate protein functions through shifts in their free energy landscapes. Ideally, new methods should be capable of shaping specific properties of these landscapes – such as multiple defined minima and the barriers between them – during the design process. There are numerous challenges with such an approach that would explicitly consider these aspects of function, including methods and informative data at sufficient scales to train models as well as characterize functional designs. Progress will also require approaches that can simultaneously quantify and optimize these multiple objectives. Such multi-objective optimization could be contrasted to – or integrated with - designs that deconstruct coupled functions to make them modular and individually tunable, such as the sense-response systems discussed above that combine separate modules for sensing and responding.

Extracting principles—As designing functional proteins beyond simpler model systems becomes possible, extracting principles becomes important. In particular, principles are needed such that *de novo* protein systems are actually tunable, modular and controllable without extensive trial-and-error or individual optimization for new contexts (such as cell type). Ideally, designs would be the result of directed and interpretable optimization (not a black box) that can systematically vary desired properties.

Since its first applications, the field of protein design has promised fundamental insights into sequence – structure – function – dynamics relationships, “learning by building”. The growing power of engineering protein components *de novo* provides different opportunities to also probe the functional principles of proteins embedded in complex interconnected biology. At the same time, these directions will also accelerate the engineering of advanced cellular functions with *de novo* components, with ultimate applications to cell therapies.

Emergent properties and advanced cellular functions with *de novo* components—The interactions and modular combinations of naturally occurring proteins lead to emergent cellular behavior that is not displayed by the individual components alone. For example, systems of proteins undergoing reversible covalent modification (e.g., phosphorylation) with opposing regulators (e.g., kinases and phosphatases) can show ultrasensitive switching, meaning that a small change in the concentration or activity of a regulator can cause a sharp change in output (modified protein)[108]. In nature, such switches are assembled into cascades for signal amplification[108]. As a second example, interlinked positive and negative regulation can control cell "states", meaning long-term, stable patterns of gene expression[109], that can be responsive to environmental signals. Already, the modularity of existing proteins can be used to reprogram advanced cellular functions[110], and machine learning can guide modular engineering[111]. The concept of composing protein systems from *de novo* designed elements should allow bottom-up design to make these advanced biological functions engineerable. This approach should allow both deconstruction and construction.

Conclusion

The field of computational *de novo* design is making a step change into a new beginning. Advances in AI applied to protein design now make many, albeit relatively simple, design goals easier and more successful. Versatile protein folds and even large protein assemblies – which already have exciting clinical applications as vaccines – can be engineered with high structural accuracy. It is increasingly possible to engineer *de novo* proteins that bind tightly to user-specified surfaces on target proteins. Applications of these *de novo* binding proteins range from probes for fundamental cell biology to therapeutic candidates. Long-standing goals of *de novo* design, such as proteins sensing new small molecule signals, design of advanced functions involving conformational changes and allostery, and engineering emergent behaviors such as ultrasensitive switching, still pose significant challenges but are coming within reach. Progress is also being made with interfacing designed systems with biology, for example to control the geometry, localization, and timing of cellular assembly processes.

Numerous exciting challenges lie ahead. Current frontiers include prediction of protein behavior beyond structure: quantitative parameters such as binding affinities, conformational dynamics, and ultimately cellular functions. Advances in deep learning will require informative data at sufficient scales to enable accurate design of these behaviors. Advanced protein functions are often composite, coupling input signals to diverse functional outputs; predictive design should hence be capable of integrating multiple objectives. Extracting principles from data is important to make desired protein properties indeed engineerable. New opportunities lie in building complex functions from the ground up. Here, *de novo* proteins could be designed *a priori* with engineering principles of tunability, controllability, and modularity. Families of such *de novo* components with tunable and controllable properties could be recombined to generate diverse behaviors. Interfacing these *de novo* systems with biological processes could enable both deconstructing cellular functions and controlling them. The rapidly evolving field of *de novo* protein design provides an exciting

environment for the creativity of scientists and engineers to address the many more unsolved than “solved” challenges at the interfaces of biological and new-to-nature functions.

Acknowledgements

I would like to thank my colleagues, especially my group past and present, for the many discussions and scientific contributions shaping this perspective on protein design. Deniz Akpinaroglu, Stephanie Crilly, and Philipp Huettemann provided insightful comments on the manuscript.

Funding:

This work was supported by a grant from the National Institutes of Health (R35 GM145236). The author is a Chan Zuckerberg Investigator.

REFERENCES

1. Regan L, and DeGrado WF (1988). Characterization of a helical protein designed from first principles. Science (New York, N.Y 241, 976–978. 10.1126/science.3043666. [PubMed: 3043666]
2. Arnold FH (2019). Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). Angew Chem Int Ed Engl 58, 14420–14426. 10.1002/anie.201907729. [PubMed: 31433107]
3. Gordley RM, Bugaj LJ, and Lim WA (2016). Modular engineering of cellular signaling proteins and networks. Current opinion in structural biology 39, 106–114. 10.1016/j.sbi.2016.06.012. [PubMed: 27423114]
4. Pan X, and Kortemme T (2021). Recent advances in de novo protein design: principles, methods, and applications. The Journal of biological chemistry, 100558. 10.1016/j.jbc.2021.100558. [PubMed: 33744284]
5. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. 10.1038/s41586-021-03819-2.
6. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science (New York, N.Y. 10.1126/science.abj8754.
7. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (New York, N.Y 379, 1123–1130. 10.1126/science.ade2574. [PubMed: 36927031]
8. Dahiyat BI, and Mayo SL (1997). De novo protein design: fully automated sequence selection. Science (New York, N.Y 278, 82–87. [PubMed: 9311930]
9. Gordon DB, Marshall SA, and Mayo SL (1999). Energy functions for protein design. Current opinion in structural biology 9, 509–513. [PubMed: 10449371]
10. Reynolds KA, Russ WP, Socolich M, and Ranganathan R (2013). Evolution-based design of proteins. Methods in enzymology 523, 213–235. 10.1016/B978-0-12-394292-0.00010-2. [PubMed: 23422432]
11. Lovelock SL, Crawshaw R, Basler S, Levy C, Baker D, Hilvert D, and Green AP (2022). The road to fully programmable protein catalysis. Nature 606, 49–58. 10.1038/s41586-022-04456-z. [PubMed: 35650353]
12. Ponder JW, and Richards FM (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of molecular biology 193, 775–791. [PubMed: 2441069]
13. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, and Baker D (2003). Design of a novel globular protein fold with atomic-level accuracy. Science (New York, N.Y 302, 1364–1368. [PubMed: 14631033]
14. Ollikainen N, Smith CA, Fraser JS, and Kortemme T (2013). Flexible backbone sampling methods to model and design protein alternative conformations. Methods in enzymology 523, 61–85. 10.1016/B978-0-12-394292-0.00004-7. [PubMed: 23422426]

15. Georgiev I, Keedy D, Richardson JS, Richardson DC, and Donald BR (2008). Algorithm for backrub motions in protein design. Bioinformatics 24, i196–204. btn169 [pii] 10.1093/bioinformatics/btn169. [PubMed: 18586714]

16. Davey JA, and Chica RA (2012). Multistate approaches in computational protein design. Protein Sci 21, 1241–1252. 10.1002/pro.2128. [PubMed: 22811394]

17. Friedland GD, and Kortemme T (2010). Designing ensembles in conformational and sequence space to characterize and engineer proteins. Current opinion in structural biology 20, 377–384. 10.1016/j.sbi.2010.02.004. [PubMed: 20303740]

18. Ferruz N, Heinzinger M, Akdel M, Goncearenco A, Naef L, and Dallago C (2023). From sequence to function through structure: Deep learning for protein design. Comput Struct Biotechnol J 21, 238–250. 10.1016/j.csbj.2022.11.014. [PubMed: 36544476]

19. Strokach A, and Kim PM (2022). Deep generative modeling for protein design. Current opinion in structural biology 72, 226–236. 10.1016/j.sbi.2021.11.008. [PubMed: 34963082]

20. Ferruz N, Schmidt S, and Hocker B (2022). ProtGPT2 is a deep unsupervised language model for protein design. Nat Commun 13, 4348. 10.1038/s41467-022-32007-7. [PubMed: 35896542]

21. Madani A, Krause B, Greene ER, Subramanian S, Mohr BP, Holton JM, Olmos JL Jr., Xiong C, Sun ZZ, Socher R, et al. (2023). Large language models generate functional protein sequences across diverse families. Nature biotechnology 41, 1099–1106. 10.1038/s41587-022-01618-2.

22. Verkuil R, Kabeli O, Du Y, Wicky BIM, Milles LF, Dauparas J, Baker D, Ovchinnikov S, Sercu T, and Rives A (2022). Language models generalize beyond natural proteins. bioRxiv, 2022.2012.2021.521521. 10.1101/2022.12.21.521521.

23. Alley EC, Khimulya G, Biswas S, AlQuraishi M, and Church GM (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 16, 1315–1322. 10.1038/s41592-019-0598-1. [PubMed: 31636460]

24. Shin JE, Riesselman AJ, Kollasch AW, McMahon C, Simon E, Sander C, Manglik A, Kruse AC, and Marks DS (2021). Protein design and variant prediction using autoregressive generative models. Nat Commun 12, 2403. 10.1038/s41467-021-22732-w. [PubMed: 33893299]

25. Hie BL, Shanker VR, Xu D, Bruun TUJ, Weidenbacher PA, Tang S, Wu W, Pak JE, and Kim PS (2023). Efficient evolution of human antibodies from general protein language models. Nature biotechnology. 10.1038/s41587-023-01763-2.

26. Anand N, Eguchi R, Mathews II, Perez CP, Derry A, Altman RB, and Huang PS (2022). Protein sequence design with a learned potential. Nat Commun 13, 746. 10.1038/s41467-022-28313-9. [PubMed: 35136054]

27. Ingraham J, Garg V, Barzilay R, and Jaakkola T (2019). Generative Models for Graph-Based Protein Design. In Wallach H, Larochelle H, Beygelzimer A, Alch\'-Buc F.d.t., Fox E, and Garnett R, eds.

28. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science (New York, N.Y 378, 49–56. 10.1126/science.add2187. [PubMed: 36108050]

29. Dieckhaus H, Brocidiacono M, Randolph N, and Kuhlman B (2023). Transfer learning to leverage larger datasets for improved prediction of protein stability changes. bioRxiv, 2023.2007.2027.550881. 10.1101/2023.07.27.550881.

30. Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, Mangan NM, Ovchinnikov S, and Rocklin GJ (2023). Mega-scale experimental analysis of protein folding stability in biology and design. Nature 620, 434–444. 10.1038/s41586-023-06328-6. [PubMed: 37468638]

31. Akpinaroglu D, Seki K, Guo A, Zhu E, Kelly MJ, and Kortemme T (2023). Structure-conditioned masked language models for protein sequence design generalize beyond the native sequence space. bioRxiv, 2023.2012.2015.571823. 10.1101/2023.12.15.571823.

32. Ollikainen N, de Jong RM, and Kortemme T (2015). Coupling Protein Side-Chain and Backbone Flexibility Improves the Re-design of Protein-Ligand Specificity. PLoS computational biology 11, e1004335. 10.1371/journal.pcbi.1004335. [PubMed: 26397464]

33. Kapp GT, Liu S, Stein A, Wong DT, Remenyi A, Yeh BJ, Fraser JS, Taunton J, Lim WA, and Kortemme T (2012). Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. Proceedings of the National Academy of Sciences of the United States of America 109, 5277–5282. 10.1073/pnas.1114487109. [PubMed: 22403064]
34. Moutevelis E, and Woolfson DN (2009). A periodic table of coiled-coil protein structures. Journal of molecular biology 385, 726–732. 10.1016/j.jmb.2008.11.028. [PubMed: 19059267]
35. Grigoryan G, and Degrado WF (2011). Probing designability via a generalized model of helical bundle geometry. Journal of molecular biology 405, 1079–1100. 10.1016/j.jmb.2010.08.058. [PubMed: 20932976]
36. Huang PS, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, and Baker D (2014). High thermodynamic stability of parametrically designed helical bundles. Science (New York, N.Y 346, 481–485. 10.1126/science.1257481. [PubMed: 25342806]
37. Jacobs TM, Williams B, Williams T, Xu X, Eletsky A, Federizon JF, Szyperski T, and Kuhlman B (2016). Design of structurally distinct proteins using strategies inspired by evolution. Science (New York, N.Y 352, 687–690. 10.1126/science.aad8036. [PubMed: 27151863]
38. Huang PS, Boyken SE, and Baker D (2016). The coming of age of de novo protein design. Nature 537, 320–327. 10.1038/nature19946. [PubMed: 27629638]
39. Sahtoe DD, Praetorius F, Courbet A, Hsia Y, Wicky BIM, Edman NI, Miller LM, Timmermans BJR, Decarreau J, Morris HM, et al. (2022). Reconfigurable asymmetric protein assemblies through implicit negative design. Science (New York, N.Y 375, eabj7662. 10.1126/science.abj7662. [PubMed: 35050655]
40. Korendovych IV, and DeGrado WF (2020). De novo protein design, a retrospective. Q Rev Biophys 53, e3. 10.1017/S0033583519000131. [PubMed: 32041676]
41. Woolfson DN (2021). A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. Journal of molecular biology 433, 167160. 10.1016/j.jmb.2021.167160. [PubMed: 34298061]
42. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, and Baker D (2012). Principles for designing ideal protein structures. Nature 491, 222–227. 10.1038/nature11600. [PubMed: 23135467]
43. Huang PS, Feldmeier K, Parmeggiani F, Velasco DAF, Hocker B, and Baker D (2016). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat Chem Biol 12, 29–34. 10.1038/nchembio.1966. [PubMed: 26595462]
44. Harteveld Z, Bonet J, Rosset S, Yang C, Sesterhenn F, and Correia BE (2022). A generic framework for hierarchical de novo protein design. Proceedings of the National Academy of Sciences of the United States of America 119, e2206111119. 10.1073/pnas.2206111119. [PubMed: 36252041]
45. Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Foight GW, Lee MY, Gagnon LA, et al. (2018). De novo design of a fluorescence-activating beta-barrel. Nature 561, 485–491. 10.1038/s41586-018-0509-0. [PubMed: 30209393]
46. Chidyausiku TM, Mendes SR, Klima JC, Nadal M, Eckhard U, Roel-Touris J, Houliston S, Guevara T, Haddox HK, Moyer A, et al. (2022). De novo design of immunoglobulin-like domains. Nat Commun 13, 5661. 10.1038/s41467-022-33004-6. [PubMed: 36192397]
47. Marcos E, Chidyausiku TM, McShan AC, Evangelidis T, Nerli S, Carter L, Nivon LG, Davis A, Oberdorfer G, Tripsianes K, et al. (2018). De novo design of a non-local beta-sheet protein with high stability and accuracy. Nat Struct Mol Biol 25, 1028–1034. 10.1038/s41594-018-0141-6. [PubMed: 30374087]
48. Khmelinskaia A, Wargacki A, and King NP (2021). Structure-based design of novel polyhedral protein nanomaterials. Curr Opin Microbiol 61, 51–57. 10.1016/j.mib.2021.03.003. [PubMed: 33784513]
49. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, and Baker D (2020). Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences of the United States of America 117, 1496–1503. 10.1073/pnas.1914677117. [PubMed: 31896580]

50. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, et al. (2021). De novo protein design by deep network hallucination. Nature 600, 547–552. 10.1038/s41586-021-04184-w. [PubMed: 34853475]
51. Wicky BIM, Milles LF, Courbet A, Ragotte RJ, Dauparas J, Kinfu E, Tipps S, Kibler RD, Baek M, DiMaio F, et al. (2022). Hallucinating symmetric protein assemblies. Science (New York, N.Y 378, 56–61. 10.1126/science.add1964. [PubMed: 36108048]
52. Anand N, and Achim T (2022). Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. arXiv:2205.15019. 10.48550/arXiv.2205.15019.
53. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, et al. (2023). De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100. 10.1038/s41586-023-06415-8. [PubMed: 37433327]
54. Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, Frappier V, Lord DM, Ng-Thow-Hing C, Van Vlack ER, et al. (2023). Illuminating protein space with a programmable generative model. Nature. 10.1038/s41586-023-06728-8.
55. Woolfson DN (2023). Understanding a protein fold: The physics, chemistry, and biology of alpha-helical coiled coils. The Journal of biological chemistry 299, 104579. 10.1016/j.jbc.2023.104579. [PubMed: 36871758]
56. Harteveld Z, Hall-Beauvais AV, Morozova I, Southern J, Goverde C, Georgeon S, Rosset S, Defferrard M, Loukas A, Vandergheynst P, et al. (2023). Exploring “dark matter” protein folds using deep learning. bioRxiv, 2023.2008.2030.555621. 10.1101/2023.08.30.555621.
57. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, et al. (2017). Global analysis of protein folding using massively parallel design, synthesis, and testing. Science (New York, N.Y 357, 168–175. 10.1126/science.aan0693. [PubMed: 28706065]
58. Pan X, Thompson MC, Zhang Y, Liu L, Fraser JS, Kelly MJS, and Kortemme T (2020). Expanding the space of protein geometries by computational design of de novo fold families. Science (New York, N.Y 369, 1132–1136. 10.1126/science.abc0881. [PubMed: 32855341]
59. Basanta B, Bick MJ, Bera AK, Norn C, Chow CM, Carter LP, Goreshnik I, Dimaio F, and Baker D (2020). An enumerative algorithm for de novo design of proteins with diverse pocket structures. Proceedings of the National Academy of Sciences of the United States of America 117, 22135–22145. 10.1073/pnas.2005412117. [PubMed: 32839327]
60. Linsky TW, Noble K, Tobin AR, Crow R, Carter L, Urbauer JL, Baker D, and Strauch EM (2022). Sampling of structure and sequence space of small protein folds. Nat Commun 13, 7151. 10.1038/s41467-022-34937-8. [PubMed: 36418330]
61. Courbet A, Hansen J, Hsia Y, Bethel N, Park YJ, Xu C, Moyer A, Boyken SE, Ueda G, Nattermann U, et al. (2022). Computational design of mechanically coupled axle-rotor protein assemblies. Science (New York, N.Y 376, 383–390. 10.1126/science.abm1183. [PubMed: 35446645]
62. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Röthlisberger D, and Baker D (2006). New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15, 2785–2794. 10.1110/ps.062353106. [PubMed: 17132862]
63. Polizzi NF, and DeGrado WF (2020). A defined structural unit enables de novo design of small-molecule-binding proteins. Science (New York, N.Y 369, 1227–1233. 10.1126/science.abb8330. [PubMed: 32883865]
64. Lucas JE, and Kortemme T (2020). New computational protein design methods for de novo small molecule binding sites. PLoS computational biology 16, e1008178. 10.1371/journal.pcbi.1008178. [PubMed: 33017412]
65. Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park JS, Jude KM, Markovic I, Kadam RU, Verschueren KHG, et al. (2022). Design of protein-binding proteins from the target structure alone. Nature 605, 551–560. 10.1038/s41586-022-04654-9. [PubMed: 35332283]
66. Gainza P, Wehrle S, Van Hall-Beauvais A, Marchand A, Scheck A, Harteveld Z, Buckley S, Ni D, Tan S, Sverrisson F, et al. (2023). De novo design of protein interactions with learned surface fingerprints. Nature 617, 176–184. 10.1038/s41586-023-05993-x. [PubMed: 37100904]

67. Wang J, Lisanza S, Juergens D, Tischer D, Watson JL, Castro KM, Ragotte R, Saragovi A, Milles LF, Baek M, et al. (2022). Scaffolding protein functional sites using deep learning. Science (New York, N.Y 377, 387–394. 10.1126/science.abn2100. [PubMed: 35862514]
68. Marchand A, Van Hall-Beauvais AK, and Correia BE (2022). Computational design of novel protein-protein interactions - An overview on methodological approaches and applications. Current opinion in structural biology 74, 102370. 10.1016/j.sbi.2022.102370. [PubMed: 35405427]
69. Silva DA, Yu S, Ulge UY, Spangler JB, Jude KM, Labao-Almeida C, Ali LR, Quijano-Rubio A, Ruterbusch M, Leung I, et al. (2019). De novo design of potent and selective mimics of IL-2 and IL-15. Nature 565, 186–191. 10.1038/s41586-018-0830-7. [PubMed: 30626941]
70. Levy S, Somasundaram L, Raj IX, Ic-Mex D, Phal A, Schmidt S, Ng WI, Mar D, Decarreau J, Moss N, et al. (2022). dCas9 fusion to computer-designed PRC2 inhibitor reveals functional TATA box in distal promoter region. Cell Rep 38, 110457. 10.1016/j.celrep.2022.110457. [PubMed: 35235780]
71. Chevalier BS, Kortemme T, Chadsey MS, Baker D, Monnat RJ, and Stoddard BL (2002). Design, activity, and structure of a highly specific artificial endonuclease. Molecular cell 10, 895–905. [PubMed: 12419232]
72. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, and Baker D (2004). Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11, 371–379. [PubMed: 15034550]
73. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, and Baker D (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science (New York, N.Y 332, 816–821. 10.1126/science.1202617. [PubMed: 21566186]
74. Goudy OJ, Nallathambi A, Kinjo T, Randolph NZ, and Kuhlman B (2023). In silico evolution of autoinhibitory domains for a PD-L1 antagonist using deep learning models. Proceedings of the National Academy of Sciences of the United States of America 120, e2307371120. 10.1073/pnas.2307371120. [PubMed: 38032933]
75. Eguchi RR, Choe CA, Parekh U, Khalek IS, Ward MD, Vithani N, Bowman GR, Jardine JG, and Huang P-S (2022). Deep Generative Design of Epitope-Specific Binding Proteins by Latent Conformation Optimization. bioRxiv, 2022.2012.2022.521698. 10.1101/2022.12.22.521698.
76. Jefferson RE, Oggier A, Fuglistaler A, Camviel N, Hijazi M, Villarreal AR, Arber C, and Barth P (2023). Computational design of dynamic receptor-peptide signaling complexes applied to chemotaxis. Nat Commun 14, 2875. 10.1038/s41467-023-38491-9. [PubMed: 37208363]
77. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, and Baker D (2013). Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501, 212–216. 10.1038/nature12443. [PubMed: 24005320]
78. Glasgow AA, Huang YM, Mandell DJ, Thompson M, Ritterson R, Loshbaugh AL, Pellegrino J, Krivacic C, Pache RA, Barlow KA, et al. (2019). Computational design of a modular protein sense-response system. Science (New York, N.Y 366, 1024–1028. 10.1126/science.aax8780. [PubMed: 31754004]
79. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR, et al. (2023). Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. bioRxiv, 2023.2010.2009.561603. 10.1101/2023.10.09.561603.
80. Ambroggio XI, and Kuhlman B (2006). Computational design of a single amino acid sequence that can switch between two distinct protein folds. Journal of the American Chemical Society 128, 1154–1161. [PubMed: 16433531]
81. Davey JA, Damry AM, Goto NK, and Chica RA (2017). Rational design of proteins that exchange on functional timescales. Nat Chem Biol 13, 1280–1285. 10.1038/nchembio.2503. [PubMed: 29058725]
82. Praetorius F, Leung PJY, Tessmer MH, Broerman A, Demakis C, Dishman AF, Pillai A, Idris A, Juergens D, Dauparas J, et al. (2023). Design of stimulus-responsive two-state hinge proteins. Science (New York, N.Y 381, 754–760. 10.1126/science.adg7731. [PubMed: 37590357]

83. Wayment-Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, Homberger M, Ovchinnikov S, Colwell L, and Kern D (2023). Predicting multiple conformations via sequence clustering and AlphaFold2. Nature. 10.1038/s41586-023-06832-9.
84. Zheng S, He J, Liu C, Shi Y, Lu Z, Feng W, Ju F, Wang J, Zhu J, Min Y, et al. (2023). Towards Predicting Equilibrium Distributions for Molecular Systems with Deep Learning. ArXiv abs/2306.05445.
85. Ramaswamy VK, Musson SC, Willcocks CG, and Degiacomi MT (2021). Deep Learning Protein Conformational Space with Convolutions and Latent Interpolations. Physical Review X 11, 011052. 10.1103/PhysRevX.11.011052.
86. Gainza-Cirauqui P, and Correia BE (2018). Computational protein design-the next generation tool to expand synthetic biology applications. Curr Opin Biotechnol 52, 145–152. 10.1016/j.copbio.2018.04.001. [PubMed: 29729544]
87. Herud-Sikimic O, Stiel AC, Kolb M, Shanmugaratnam S, Berendzen KW, Feldhaus C, Hocker B, and Jurgens G (2021). A biosensor for the direct visualization of auxin. Nature 592, 768–772. 10.1038/s41586-021-03425-2. [PubMed: 33828298]
88. Chen KM, Keri D, and Barth P (2020). Computational design of G Protein-Coupled Receptor allosteric signal transductions. Nat Chem Biol 16, 77–86. 10.1038/s41589-019-0407-2. [PubMed: 31792443]
89. Paradis JS, Feng X, Murat B, Jefferson RE, Sokrat B, Szpakowska M, Hogue M, Bergkamp ND, Heydenreich FM, Smit MJ, et al. (2022). Computationally designed GPCR quaternary structures bias signaling pathway activation. Nat Commun 13, 6826. 10.1038/s41467-022-34382-7. [PubMed: 36369272]
90. Foight GW, Wang Z, Wei CT, Jr Greisen P, Warner KM, Cunningham-Bryant D, Park K, Brunette TJ, Sheffler W, Baker D, and Maly DJ (2019). Multi-input chemical control of protein dimerization for programming graded cellular responses. Nature biotechnology 37, 1209–1216. 10.1038/s41587-019-0242-8.
91. Shui S, Gainza P, Scheller L, Yang C, Kurumida Y, Rosset S, Georgeon S, Di Roberto RB, Castellanos-Rueda R, Reddy ST, and Correia BE (2021). A rational blueprint for the design of chemically-controlled protein switches. Nat Commun 12, 5754. 10.1038/s41467-021-25735-9. [PubMed: 34599176]
92. Kretschmer S, and Kortemme T (2022). Advances in the Computational Design of Small-Molecule-Controlled Protein-Based Circuits for Synthetic Biology. Proc IEEE Inst Electr Electron Eng 110, 659–674. 10.1109/JPROC.2022.3157898. [PubMed: 36531560]
93. Shui S, Scheller L, and Correia BE (2023). Protein-based bandpass filters for controlling cellular signaling with chemical inputs. Nat Chem Biol. 10.1038/s41589-023-01463-7.
94. Steiner PJ, Swift SD, Bedewitz M, Wheeldon I, Cutler SR, Nusinow DA, and Whitehead TA (2023). A Closed Form Model for Molecular Ratchet-Type Chemically Induced Dimerization Modules. Biochemistry 62, 281–291. 10.1021/acs.biochem.2c00172. [PubMed: 35675717]
95. Ng AH, Nguyen TH, Gomez-Schiavon M, Dods G, Langan RA, Boyken SE, Samson JA, Waldburger LM, Dueber JE, Baker D, and El-Samad H (2019). Modular and tunable biological feedback control using a de novo protein switch. Nature. 10.1038/s41586-019-1425-7.
96. Langan RA, Boyken SE, Ng AH, Samson JA, Dods G, Westbrook AM, Nguyen TH, Lajoie MJ, Chen Z, Berger S, et al. (2019). De novo design of bioactive protein switches. Nature 572, 205–210. 10.1038/s41586-019-1432-8. [PubMed: 31341284]
97. Lajoie MJ, Boyken SE, Salter AI, Bruffey J, Rajan A, Langan RA, Olshefsky A, Muhunthan V, Bick MJ, Gewe M, et al. (2020). Designed protein logic to target cells with precise combinations of surface antigens. Science (New York, N.Y 369, 1637–1643. 10.1126/science.aba6527. [PubMed: 32820060]
98. Chen Z, Kibler RD, Hunt A, Busch F, Pearl J, Jia M, VanAernum ZL, Wicky BIM, Dods G, Liao H, et al. (2020). De novo design of protein logic gates. Science (New York, N.Y 368, 78–84. 10.1126/science.aay2790. [PubMed: 32241946]
99. Fink T, Lonzaric J, Praznik A, Plaper T, Merljak E, Leben K, Jerala N, Lebar T, Strmsek Z, Lapenta F, et al. (2019). Design of fast proteolysis-based signaling and logic circuits in mammalian cells. Nat Chem Biol 15, 115–122. 10.1038/s41589-018-0181-6. [PubMed: 30531965]

100. Hilditch AT, Romanyuk A, Cross SJ, Obexer R, McManus JJ, and Woolfson DN (2023). Assembling membraneless organelles from de novo designed proteins. Nat Chem. 10.1038/s41557-023-01321-y.

101. Mout R, Bretherton RC, Decarreau J, Lee S, Edman NI, Ahlrichs M, Hsia Y, Sahtoe DD, Ueda G, Gregorio N, et al. (2023). De novo design of modular protein hydrogels with programmable intra- and extracellular viscoelasticity. bioRxiv, 2023.2006.2002.543449. 10.1101/2023.06.02.543449.

102. Elazar A, Chandler NJ, Davey AS, Weinstein JY, Nguyen JV, Trenker R, Cross RS, Jenkins MR, Call MJ, Call ME, and Fleishman SJ (2022). De novo-designed transmembrane domains tune engineered receptor functions. eLife 11. 10.7554/eLife.75660.

103. Huang B, Abedi M, Ahn G, Coventry B, Sappington I, Wang R, Schlichthaerle T, Zhang JZ, Wang Y, Goreshnik I, et al. (2023). Designed Endocytosis-Triggering Proteins mediate Targeted Degradation. bioRxiv, 2023.2008.2019.553321. 10.1101/2023.08.19.553321.

104. Rhys GG, Cross JA, Dawson WM, Thompson HF, Shanmugaratnam S, Savery NJ, Dodding MP, Hocker B, and Woolfson DN (2022). De novo designed peptides for cellular delivery and subcellular localisation. Nat Chem Biol 18, 999–1004. 10.1038/s41589-022-01076-6. [PubMed: 35836017]

105. Watson JL, Kruger LK, Ben-Sasson AJ, Bittleston A, Shahbazi MN, Planelles-Herrero VJ, Chambers JE, Manton JD, Baker D, and Derivery E (2023). Synthetic Par polarity induces cytoskeleton asymmetry in unpolarized mammalian cells. Cell 186, 4710–4727 e4735. 10.1016/j.cell.2023.08.034. [PubMed: 37774705]

106. Edman NI, Redler RL, Phal A, Schlichthaerle T, Srivatsan SR, Etemadi A, An S, Favor A, Ehnes D, Li Z, et al. (2023). Modulation of FGF pathway signaling and vascular differentiation using designed oligomeric assemblies. bioRxiv. 10.1101/2023.03.14.532666.

107. Markin CJ, Mokhtari DA, Sunden F, Appel MJ, Akiva E, Longwell SA, Sabatti C, Herschlag D, and Fordyce PM (2021). Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. Science (New York, N.Y 373. 10.1126/science.abf8761.

108. Goldbeter A, and Koshland DE Jr. (1981). An amplified sensitivity arising from covalent modification in biological systems. Proceedings of the National Academy of Sciences of the United States of America 78, 6840–6844. 10.1073/pnas.78.11.6840. [PubMed: 6947258]

109. Zhu R, Del Rio-Salgado JM, Garcia-Ojalvo J, and Elowitz MB (2022). Synthetic multistability in mammalian cells. Science (New York, N.Y 375, eabg9765. 10.1126/science.abg9765. [PubMed: 35050677]

110. Lim WA (2022). The emerging era of cell engineering: Harnessing the modularity of cells to program complex biological function. Science (New York, N.Y 378, 848–852. 10.1126/science.add9665. [PubMed: 36423287]

111. Daniels KG, Wang S, Simic MS, Bhargava HK, Capponi S, Tonai Y, Yu W, Bianco S, and Lim WA (2022). Decoding CAR T cell phenotype using combinatorial signaling motif libraries and machine learning. Science (New York, N.Y 378, 1194–1200. 10.1126/science.abq0225. [PubMed: 36480602]

112. Ovchinnikov S, and Huang PS (2021). Structure-based protein design with deep learning. Curr Opin Chem Biol 65, 136–144. 10.1016/j.cbpa.2021.08.004. [PubMed: 34547592]

113. Fink T, and Jerala R (2022). Designed protease-based signaling networks. Curr Opin Chem Biol 68, 102146. 10.1016/j.cbpa.2022.102146. [PubMed: 35430555]

114. Alberstein RG, Guo AB, and Kortemme T (2022). Design principles of protein switches. Current opinion in structural biology 72, 71–78. 10.1016/j.sbi.2021.08.004. [PubMed: 34537489]

115. Vorobieva AA (2021). Principles and Methods in Computational Membrane Protein Design. Journal of molecular biology 433, 167154. 10.1016/j.jmb.2021.167154. [PubMed: 34271008]

116. Zhu J, and Lu P (2022). Computational design of transmembrane proteins. Current opinion in structural biology 74, 102381. 10.1016/j.sbi.2022.102381. [PubMed: 35537282]

117. Boyken SE, Chen Z, Groves B, Langan RA, Oberdorfer G, Ford A, Gilmore JM, Xu C, DiMaio F, Pereira JH, et al. (2016). De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. Science (New York, N.Y 352, 680–687. 10.1126/science.aad8865. [PubMed: 27151862]

118. Chen Z, Boyken SE, Jia M, Busch F, Flores-Solis D, Bick MJ, Lu P, VanAernum ZL, Sahasrabuddhe A, Langan RA, et al. (2019). Programmable design of orthogonal protein heterodimers. Nature 565, 106–111. 10.1038/s41586-018-0802-y. [PubMed: 30568301]

119. Hoersch D, Roh SH, Chiu W, and Kortemme T (2013). Reprogramming an ATP-driven protein machine into a light-gated nanocage. Nature nanotechnology 8, 928–932. 10.1038/nnano.2013.242.

120. Marchand A, Bonati L, Shui S, Scheller L, Gainza P, Rosset S, Georgeon S, Tang L, and Correia BE (2023). Rational Design of Chemically Controlled Antibodies and Protein Therapeutics. ACS Chem Biol 18, 1259–1265. 10.1021/acschembio.3c00012. [PubMed: 37252896]

121. Quijano-Rubio A, Yeh HW, Park J, Lee H, Langan RA, Boyken SE, Lajoie MJ, Cao L, Chow CM, Miranda MC, et al. (2021). De novo design of modular and tunable protein biosensors. Nature 591, 482–487. 10.1038/s41586-021-03258-z. [PubMed: 33503651]

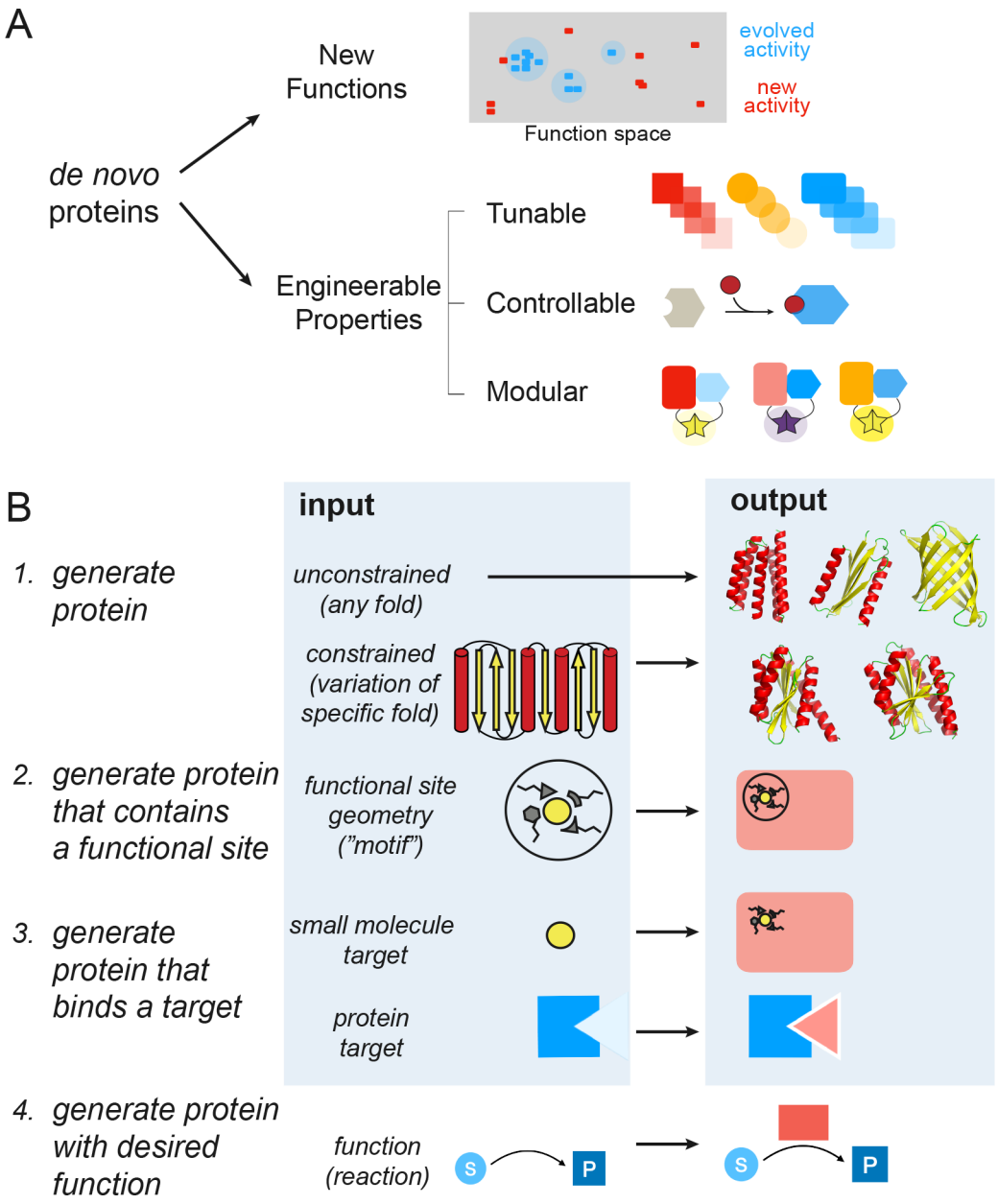

Figure 1. *De novo* protein design in the age of AI.

(**A**) Designing proteins *de novo* (from scratch, without starting from a natural protein) can explore new structures and functions, and design proteins *a priori* with engineering principles in mind: Proteins could be designed to be tunable in their quantitative properties (rates, affinities, etc.), controllable by arbitrary inputs, and modular such that protein elements can be linked together for diverse input/output behaviors. (**B**) Advances in AI change the process of *de novo* protein design. User-defined goals (left) and inputs (middle) are used to generate proteins with new structures and functions (right). Categories 1–4 depict increasingly straightforward prompts leading to increasingly complex design outputs. Blue shading indicates design goals with experimentally validated examples. B-1: AI-based methods to design new protein structures can be unconstrained (generating diverse protein

folds; alpha-helices shown in red and beta-strands in yellow) or constrained to diversify a particular fold. B-2: Most current methods to design function specify a “motif” with defined residue positions and orientations in a functional site. In a second step, a protein is generated *de novo* that surrounds and stabilizes the precise functional site geometry. This process is called “motif scaffolding”. B-3: Advances in AI-based methods are in development that only define the target, and the design method generates a predicted binder. B-4: Starting from a target function (for example converting substrate S to product P), an AI method could generate a protein with the requirements for that function. Currently, protein language models trained on specific protein families or large experimental datasets can generate new sequences with functions similar to those in the training set.

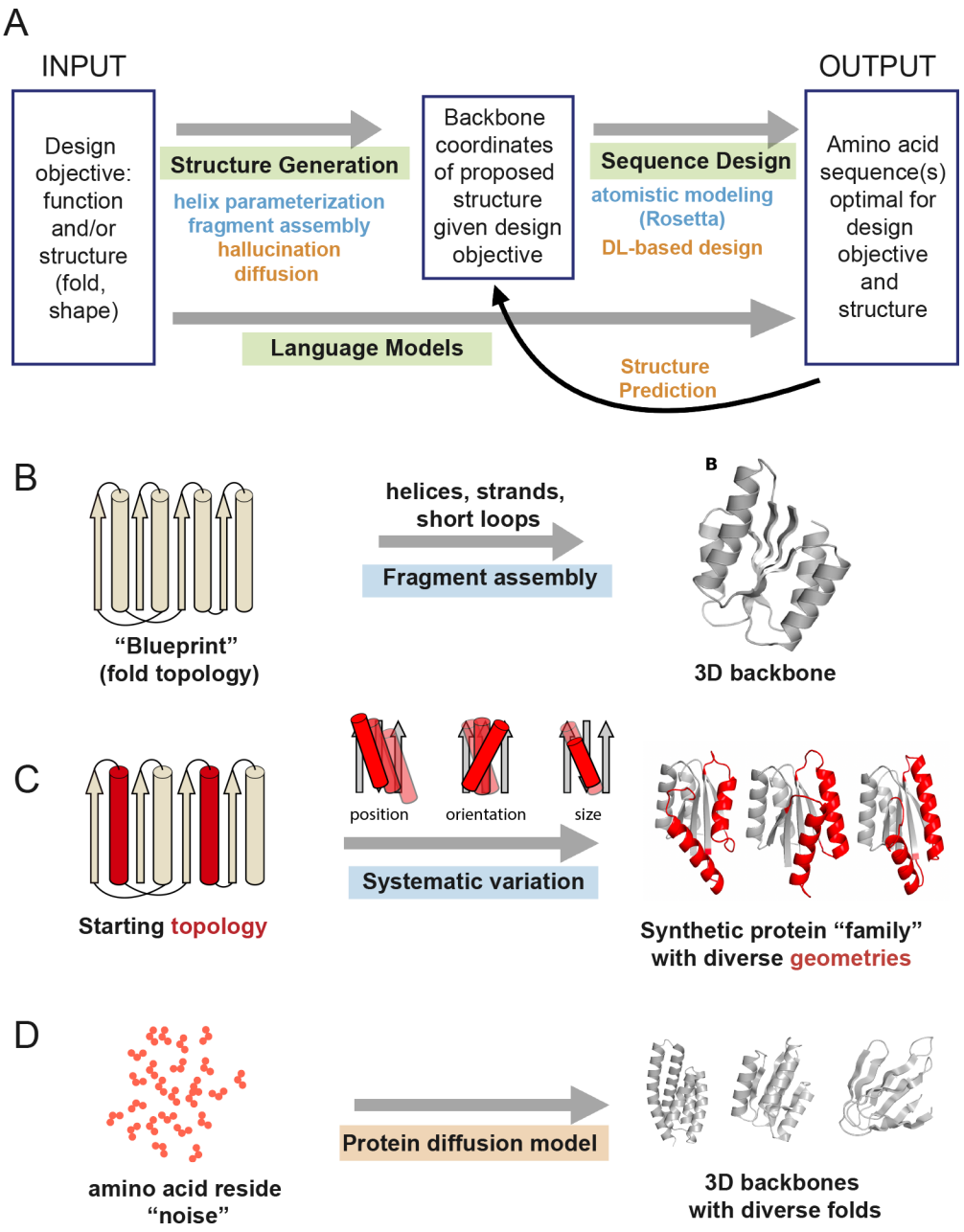

Figure 2. Protein design concepts and approaches.

(**A**) *De novo* protein design is formulated as an optimization problem: Given a design objective (a protein with a desired shape and function), find one or more amino acid sequences that have that structure and function. Most design methods divide the process into two steps: First, a structure containing only the polypeptide backbone is generated, and then a sequence is designed for that backbone. For each step, design methods that use atomistic modeling (blue) or AI-based approaches (orange) are indicated. (**B**) Classical design methods use a "blueprint" defining a protein fold topology (identity and order of secondary structure elements) and then assembles a 3-dimensional backbone from ideal helix, strand, and loop peptide fragments. (**C**) Backbone generation methods can systematically sample

geometries (positions, orientations and sizes of secondary structure elements with varied connecting loops) within a given fold. These methods generate synthetic fold families that, just like evolved protein families, can be optimized for diverse functions. (**D**) A recent AI-based method, protein diffusion, generates protein backbones through a denoising process from random backbone starting coordinates. This method generates diverse protein folds without having to pre-specify a topology as input.

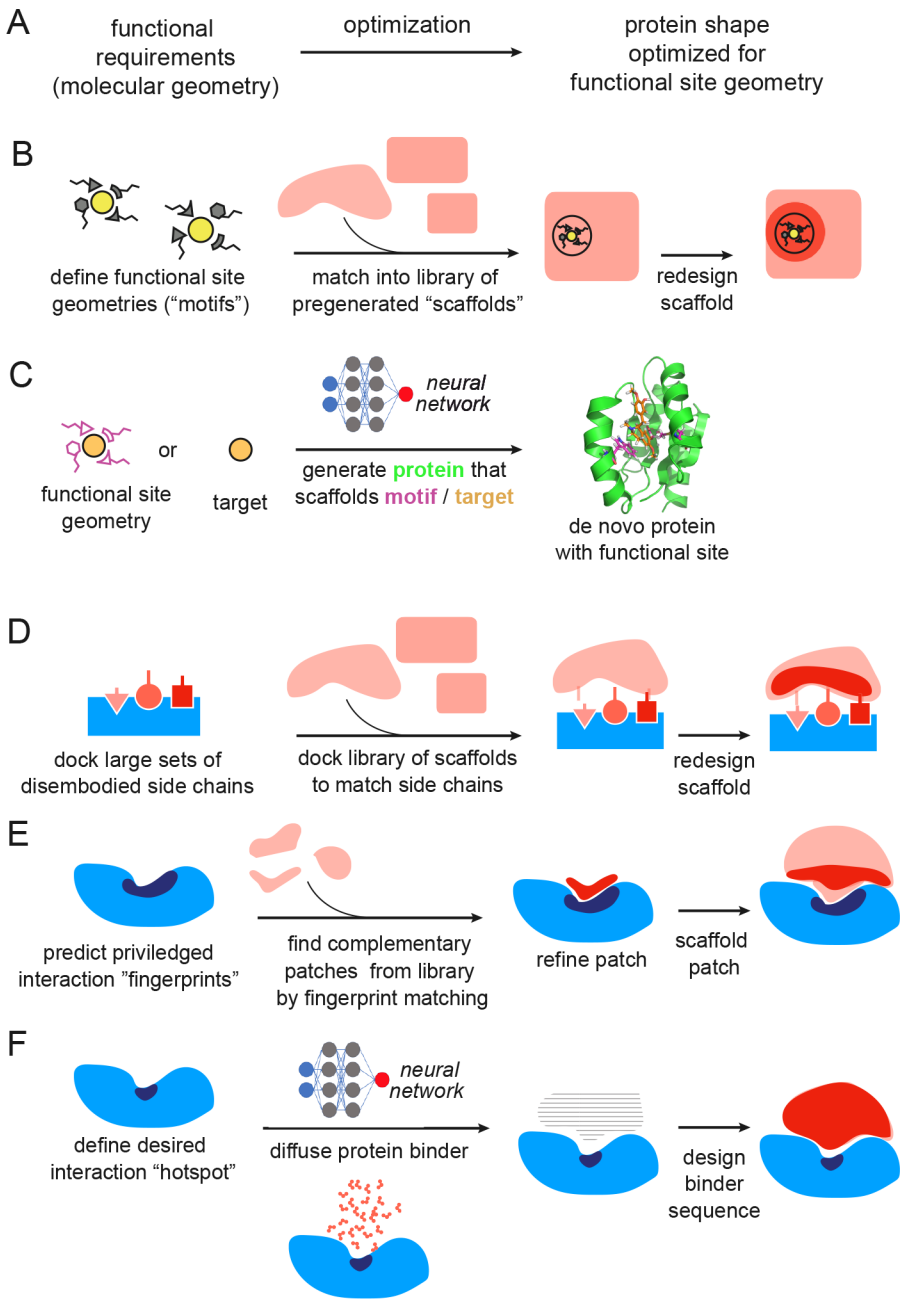

Figure 3. *De novo* design of molecular functions.

(**A**) General approach to design molecular functions. (**B-C**) Design of proteins binding to small molecules, using classical design methods (**B**) that place target binding sites into pre-generated protein scaffolds, or AI-based approaches (**C**) that generate new protein backbones around a binding site motif or target. (**D-F**) Design of proteins binding to target proteins (blue shapes). Regions that are optimized by sequence design are shown as dark red shape. (**D**) Rotamer interaction field approach[65]. Specific interactions with a target protein surface are identified through docking of disembodied side chains, yielding an interaction field into which pre-existing scaffolds are docked and optimized. (**E**) Fingerprint approach[66]. Interaction sites on the target are identified by predicting interaction fingerprints

using the MaSIF deep-learning method, followed by identification of complementary fingerprints from a library of >400 Million patches. Matching patches are then scaffolded into *de novo* proteins and optimized. (**F**) Diffusion approach[53]. AI-based protein diffusion is used to generate a binding protein with shape complementarity to a pre-specified hotspot on the target. A second step assigns a sequence to the diffused binder backbone.

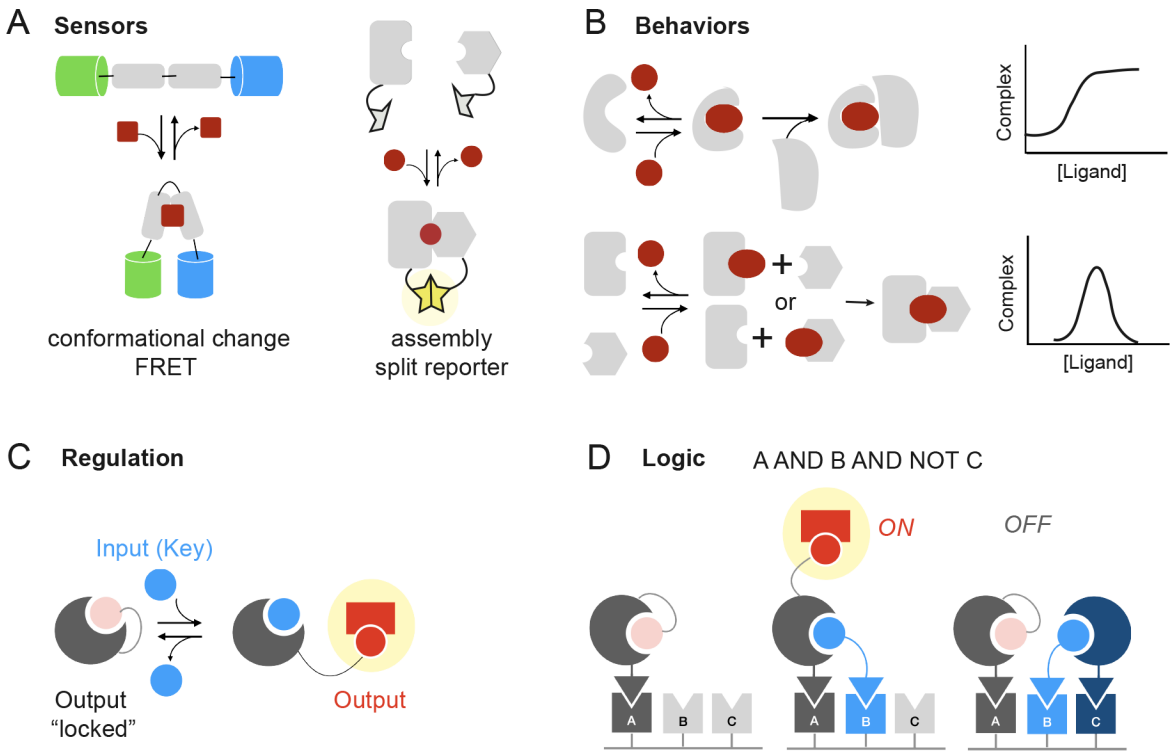

Figure 4. *De novo* design to control cellular functions.

(A) Computational design of small-molecule sensors that couple auxin ligand binding to conformational change and fluorescence energy transfer (FRET)[87] (left) or metabolite-induced protein-protein dimerization to split reporter complementation[78] (right). (**B**) Different quantitative behaviors for CID systems. Top: "ratchet" mechanism, where ligand binding leads to a conformational change in one protein that creates a composite binding interface for the second protein. Bottom: "molecular glue" mechanism where the small molecule can bind either partner. This mechanism can lead to "bandpass filter" behavior where complex formation is low at high ligand concentrations because each of the two protein partners are bound by a different ligand molecule. (**C**) Mechanism of the *de novo* designed LOCKR system, where an output element is buried but can be displaced by a competing key element, leading to an output. (**D**) Application of the Co-LOCKR system to perform logic operations based on the composition of receptors present on the cell surface[97].

Table 1.

Recent computational protein design reviews with title, short summary, and reference.

Protein design concepts and progression of the field		
Recent advances in de novo protein design: principles, methods, and applications.	State of protein design before broader adoption of AI-based methods: generation of backbone structures, sequence optimization, design energy functions, and design of molecular functions.	**4**
De novo Protein Design, a Retrospective	Evolution of the field of *de novo* protein design, with focus on physicochemical principles, functional helical bundles, membrane proteins, and protein assemblies.	**40**
A Brief History of de novo Protein Design: Minimal, Rational, and Computational	Progress in protein design illustrated through a timeline of *de novo* protein structures solved to atomic resolution.	**41**
Understanding a protein fold: The physics, chemistry, and biology of alpha-helical coiled coils	Progress in understanding and engineering alpha-helical coiled coils including design principles, biological functions, and applications of coiled coils in synthetic biology.	**55**
Machine / deep learning		
Structure-Based Protein Design with Deep Learning	Outline of deep learning approaches to protein design and comparison to prior design methods.	**112**
Deep Generative Modeling for Protein Design	Comparison of 5 classes of generative models used for protein design.	**19**
From Sequence to Function through Structure: Deep Learning for Protein Design	Summary and comprehensive tables of recent deep learning methods for (i) fixed backbone sequence design, (ii) structure generation, (iii) sequence generation, and (iv) concomitant design of sequence and structure.	**18**
Protein-protein interactions		
Computational Design of Novel ProteinProtein Interactions - An Overview on Methodological Approaches and Applications	Methods and successful cases of designing protein-protein interactions using (i) template-based approaches (utilizing known protein-protein interactions) and (ii) *de novo* design.	**68**
Applications to biological engineering		
Computational Protein Design-the next Generation Tool to Expand Synthetic Biology Applications	Summary of computational designs shown to modulate activities in cells, including enzymes, protein specificity engineering, cellular pathway control, and higher-order protein assemblies.	**86**
Advances in the Computational Design of Small-Molecule-Controlled Protein-Based Circuits for Synthetic Biology	Computational approaches to designing protein-based sensors for small-molecule inputs coupled to functional outputs in cells.	**92**
Designed protease-based signaling networks	Summary of approaches that have engineered protease-based synthetic circuits for cellular regulation.	**113**
Design of protein switches		
Design Principles of Protein Switches	Applications of switch design inspired by naturally occurring protein switches, and challenges with designing them *de novo*.	**114**
Membrane proteins		
Principles and Methods in Computational Membrane Protein Design	Overview of innovations in the generation of new membrane protein structures and functions.	**115**
Computational Design of Transmembrane Proteins	Principles for transmembrane protein design and successful examples.	**116**
Enzymes		
The Road to Fully Programmable Protein Catalysis	Key developments and opportunities in the challenging field of enzyme design.	**11**

Table 2.

Protein systems engineered to be tunable, controllable, and composable, with publication title, short summary, and reference.

Families of components with tunable properties		
Expanding the space of protein geometries by computational design of de novo fold families	*De novo* protein fold families with finely tunable shapes through systematic variation of helical elements.	**58**
An enumerative algorithm for de novo design of proteins with diverse pocket structures	Families of *de novo* NTF2 fold proteins with pockets tunable for ligand binding.	**59**
De novo design of protein homo-oligomers with modular hydrogen-bond networkmediated specificity	Alpha-helical homo-oligomers with diverse interaction specificity determined by central hydrogen-bond networks.	**117**
Programmable design of orthogonal protein heterodimers	Orthogonal 4-helix protein heterodimers of two helical hairpins, with interaction specificity determined by hydrogen-bond networks.	**118**
De novo design of bioactive protein switches	Orthogonal LOCKR systems that function *in vitro*, in yeast and in mammalian cells.	**96**
Reconfigurable asymmetric protein assemblies through implicit negative design	Families of beta sheet-mediated heterodimers with diverse on- and off-rates.	**39**
Controllability		
Reprogramming an ATP-driven protein machine into a light-gated nanocage	Generalizable strategy to control reversible shape changes of a protein nanocage through light-triggered conformational switching of a covalently attached azobenzene linker.	**119**
Computational design of a modular protein sense-response system	Control of protein-protein assembly through *de novo* design of a small molecule binding site into a protein-protein interface.	**78**
A rational blueprint for the design of chemically-controlled protein switches	Computational protein design strategy to repurpose drug-inhibited protein-protein interactions as OFF- and ON-switches.	**91**
Rational Design of Chemically Controlled Antibodies and Protein Therapeutics	Design and application of small-molecule-controlled switchable protein therapeutics using an engineered OFF-switch system[91] (above).	**120**
Designed protein logic to target cells with precise combinations of surface antigens	Application of the LOCKR systems[96] (above) as colocalization- dependent protein switches (Co-LOCKR) that can perform Boolean logic operations on the cell surface.	**97**
Modularity		
Computational design of a modular protein sense-response system	A *de novo* designed chemically-induced heterodimerization system[78] (above) can be linked to diverse modular split response systems.	**78**
Reconfigurable asymmetric protein assemblies through implicit negative design	Tunable beta-sheet heterodimers[39] (above) can be assembled into complexes with up to 6 different components.	**39**
Modular and Tunable Biological Feedback Control Using a de Novo Protein Switch	The LOCKR system[96] (above) can be modularly recombined and rationally tuned to implement feedback control of endogenous signaling pathways and synthetic gene circuits.	**95**
De novo design of modular and tunable protein biosensors	The LOCKR system[96] (above) can be adapted into modular biosensors for diverse proteins.	**121**