Published in final edited form as:

Nat Biotechnol. 2024 February ; 42(2): 203–215. doi:10.1038/s41587-024-02133-2.

Sparks of function by de novo protein design

Alexander E. Chu[1,2,3], **Tianyu Lu**[2], **Po-Ssu Huang**[1,2,✉]

[1]Biophysics Program, Stanford University, Palo Alto, CA, USA.

[2]Department of Bioengineering, Stanford University, Palo Alto, CA, USA.

[3]Present address: Google DeepMind, London, UK.

Information in proteins flows from sequence to structure to function, with each step causally driven by the preceding one. Protein design is founded on inverting this process: specify a desired function, design a structure executing this function, and find a sequence that folds into this structure. This ‘central dogma’ underlies nearly all de novo protein-design efforts. Our ability to accomplish these tasks depends on our understanding of protein folding and function and our ability to capture this understanding in computational methods. In recent years, deep learning-derived approaches for efficient and accurate structure modeling and enrichment of successful designs have enabled progression beyond the design of protein structures and towards the design of functional proteins. We examine these advances in the broader context of classical de novo protein design and consider implications for future challenges to come, including fundamental capabilities such as sequence and structure co-design and conformational control considering flexibility, and functional objectives such as antibody and enzyme design.

De novo protein design was born out of a desire to reduce the complexity of protein folding down to basic physical principles. It was hypothesized that, with sufficient understanding of the rules governing protein folding, it might be possible to create new proteins from scratch[1,2]. In time, this hypothesis has proven true. The guiding physical principles of protein design are simple, but the process of applying these principles leads to vastly diverse structural outcomes, unlocking a new era of functional protein design[3]. For many problems in protein design, de novo design has become more effective than computationally manipulating or adapting native protein structures to achieve a desired function[4].

✉**Correspondence and requests for materials** should be addressed to Po-Ssu Huang. possu@stanford.edu.

Author contributions

Planning, figure production and writing of the Review: all authors.

Reference curation: A.E.C. and T.L. Supplementary tables: A.E.C. and T.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-024-02133-2.

Peer review information *Nature Biotechnology* thanks Philip Kim and Kevin Yang for their contribution to the peer review of this work.

Traditionally, protein structure and its interaction with sequence are understood in an energetic and biophysical sense: what are the three-dimensional interactions that amino acid residues make with each other? How do they stabilize a particular conformation of the protein chain or an interaction with a ligand or substrate? The ability to capture the diverse behaviors of proteins with a set of atomic-level physical equations is attractive, providing an interpretable view of the forces that sustain a structure. Indeed, the earliest protein-design methods used this approach successfully to define structures of new proteins and resample side chains for new sequences[5–10].

However, the space of all possible protein conformations and sequences is far greater than can be explored exhaustively in the timescales of protein folding or evolution or any kind of computational or experimental sampling scheme[11,12]. Yet somehow, through billions of years of evolution, nature has managed to produce a small set of proteins. For scientists who wish to solve problems on shorter timescales, drawing on data from nature's 'answer key' has been a highly effective strategy. Since the first design of a new protein fold by assembling fragments of natural proteins[13], protein data available in the Protein Data Bank (PDB) have grown rapidly. This has enabled an increasing role for incorporating data in protein design through tools such as structural fragment libraries, scoring functions regressed to data, sequence and rotamer statistics[14–17], eventually leading to the design of protein structures with atomic accuracy[3].

As methods for de novo design matured, it became relevant to also consider protein function. Could protein structures and sequences not only be designed from scratch but also accomplish desired functions? In contrast to redesign of natural proteins, de novo approaches offer complete control over the structure and sequence, whereas natural proteins are often marginally stable and marginally functional. It can be hard to predict when an engineering change will result in an unfoldable protein. In recent years, our ability to design functional proteins has seen a step change, as fast, performant structure-design models combined with precise validation of designed sequences with AlphaFold have led to a new age of functional design in which proteins are designed from scratch to conform to functional motifs, rather than altered from existing proteins (whether de novo or natural) to support these motifs. This has unlocked several applications, including supra-molecular assemblies, transmembrane pores and protein, ligand and metal binders[18–21].

In this Review, we examine each of the three pillars of the 'central dogma' of de novo protein design: (1) how functional goals can be mapped to structural motifs accomplishing these goals, (2) how we can control and design protein structure, especially in response to these motifs, and (3) how sequence is sampled so that the designed structure is attained and functional roles are fulfilled by side chains (Fig. 1). For each pillar, we discuss the insights and strategies that have arisen to enable more accurate protein design and survey the key methods that have demonstrated our improved capacity to generate functional proteins. We explore the potential of new approaches that expand beyond the current paradigm, including alternate ways to model structure and sequence and modeling of conformational dynamics and heterogeneity. We conclude by discussing remaining challenges for functional design and give an outlook for the field. Methods are summarized in Supplementary Table 1; in addition to works cited below, see refs. 22–51 therein.

Deriving structure from function

De novo design of a functional protein begins with identifying the features needed to accomplish the intended function. Examples of common objectives include designing proteins to engage immune cells, creating binders for drugs, nucleic acids or other proteins, stabilizing the transition state of a reaction for new enzymes and developing ion-specific transmembrane channels. Regardless of the application, the approaches are built on the principles of energetic stabilization and shape complementarity[1,52]. In earlier de novo design efforts, the design of any foldable protein was already considered a major achievement, and efforts to attain function centered on introducing changes to these scaffolds to accommodate a functional motif (in a minimal way)[2,53]. With the rise of increasingly powerful design methods, specifying the functional motif first and then searching for protein scaffolds that are consistent with this motif has become the more common path.

In many cases, the relevant functional motifs can be extracted directly from natural proteins and scaffolded as part of a de novo protein structure. This strategy has been deployed to scaffold antigen epitopes on the surface of designed immunogens[54]. Other successes with this approach include scaffolding peptide-binding motifs, metal-binding sites and ligand-binding motifs to accomplish the relevant functional task[55,56]. These motifs can also be extracted from nature to support a designed function, such as the placement of positively charged residues near the membrane–solvent interface in the case of designing transmembrane channels[19,57,58].

This approach requires known solutions from already functional proteins for the problem of interest. More general approaches to devise yet unknown functional motifs require breaking down the interaction to basic chemical elements and handling the possible combinations and arrangements of these elements accurately. One class of methods solves this problem by considering the chemical properties of the target and enumerating the interactions that a protein might use to bind to the target (Fig. 2)[59,60]. They can also be culled directly from the PDB, relying on statistical enrichment to capture the most effective interactions and perhaps average out noisier information such as side chain flexibility[61,62]. For these side chain-focused methods, choosing small, fine-grained chemical groups (such as amides or carbonyls) increases the number of unique examples and enables generalization to more complex motifs. This interaction field approach is generalizable to arbitrary binding interactions and has been successfully applied to design de novo binders against conformer-specific small-molecule ligands[59,61], miniprotein binders and ultra-high-affinity de novo binders to receptors[60], monobody binders to nerve toxins[62] and binding to nucleic acids[63].

Other approaches seek a higher level of abstraction by capturing features of functional interfaces with machine learning. For protein–protein interfaces, machine-learned representations of a surface can be used to propose the binding counterpart. Embeddings of protein surfaces can be learned that capture general biophysical and biochemical properties of an interface region as well as additional information that may be encoded in subtle variations in the sequence but is difficult to explain with energy functions or visual inspection. The patch embeddings from a target protein can then be inverted and mapped to

sets of favorably interacting motifs for scaffolding into a designed protein; this approach has been shown to yield high-affinity protein binders to diverse targets (Fig. 2)[64,65].

For some binding interactions, extracting interface features can be bypassed entirely if a model can be trained on relevant data, which allows generalizing to new, unseen examples. This zero-shot transfer learning approach allows the model to reuse learning from data-rich regimes on problems that are data poor. For example, there are far fewer protein complexes than monomers in the PDB, but models that are trained on monomers can still learn about protein complexes through the features shared by both types of data: the fundamental forces governing interactions between amino acids. Even if the models are not explicitly trained to capture the physics, similar underlying patterns appear in both monomers and complexes, allowing for improved performance on the latter. In practice, models can be further finetuned on scarcer task-specific data beyond relying solely on model generalization. This concept has been implemented to generate high-affinity binders for multiple targets[20,66] and has also been extended to small-molecule ligands[21].

Designing structure from scratch

With a functional motif defined, devising a protein structure to satisfy the constraints posed by it is one of the most challenging aspects of protein design. We previously reviewed conventional backbone-design methods[3] and suggested that the space of designable sequences might remain largely unexplored. De novo protein design enables this exploration by breaking down structure design into the hierarchical components of topology, which defines the sequence and arrangement of secondary-structure elements, and syntax, which defines the lengths of these elements[67,68]. In conventional protein design, these definitions are captured in a blueprint, which can be implemented by fragment-assembly routines[14,69].

Designing protein structures in this conventional fashion still offers the most interpretable way to model protein structures. For example, designs incorporating key structural insights have refined our ability to control β-barrel-forming structures, which are important in enzymes and membrane protein applications[19,59,70]. Since the principles for building triosephosphate isomerase (TIM) barrels were first established[71], altering the central β-barrel to have an ovoid (rather than circular) shape has been a major goal, because it is more suitable for incorporating small-molecule-binding sites. This proved to be difficult by conventional protein engineering, but revisiting the basic topology revealed that an oval-shaped TIM barrel is the result of sliding two half-circular barrels along the tilted β-strands (Fig. 3a)[68]. In developing de novo β-barrel proteins into membrane proteins, Vorobieva et al. came to the important insight that destabilizing elements of a designed structure allow the peptide chain to be inserted into the membrane. The same guiding principles for building β-barrels also yielded a proof-of-principle stereoselective retro-aldolase[72]. Few other approaches can reveal the inner workings of proteins to this depth. However, despite having complete control over the construction of a structure to the individual residue level, adapting these designs for function can be difficult due to the stringent conformity to idealized building blocks. Natural proteins tolerate, even require, non-ideal structures to achieve complex function, and this inspired efforts to find more sophisticated processes to accomplish de novo protein design.

Our ability to manipulate protein structure in response to functional constraints has seen enormous change with the application of deep learning[73]. In an approach similar in spirit to the original energy landscape-based perspective on de novo design, learned and statistical potentials can be used in place of physics-based potentials to guide structure search, enabling a similar ability to produce new-to-nature structures and topologies[74]. The advent of highly accurate protein structure prediction using the AlphaFold system and the subsequent development of trRosetta and RoseTTAFold[75–79] opened new ways to generate proteins. By learning to map the distribution of protein sequences to the distribution of structures, these methods appeared to encode information about both within a single differentiable network. Efforts to tease out what these predictive models learned about structure gave rise to the class of hallucination-based approaches, which explored various ways to invert structure-prediction networks by optimizing and resampling the sequence inputs until they produced realistic output structures[56,80]. A similar approach found that a masking objective applied during RoseTTAFold training could be extended to do ‘inpainting’, that is, completing missing regions of a partially masked structure[55,81]. These approaches yield de novo proteins in a mostly automated way, without requiring the intense structural scrutiny and large-scale sampling devoted to previous design efforts. This enabled searching protein structure space broadly and quickly for solutions to design constraints, leading to successful scaffolding of various functional motifs and inputs in new de novo proteins.

In a parallel approach, deep generative modeling emerged as a powerful strategy for efficient sampling from high-dimensional distributions for which we have plentiful data, such as images and text[82–86]. These models learn to approximate a mapping from a distribution that is easy to sample from, such as a Gaussian distribution, to a data distribution of interest. This method can also be applied to protein design and provides a more natural way to generate protein structures by construction[62,87] without having to hack the inputs to a structure-prediction network. An important advance in generative protein design occurred with the rise of diffusion-based generative models, which attain high sample quality while providing more stable training and better diversity than other types of generative models[88–93]. These models benefit from an iterative generation mechanism that begins with white noise and denoises coarse features first before filling in fine details, rather than attempting to synthesize the full atomic structure in one shot. This inductive bias, or learning architecture, aligns well with the hierarchical nature of protein structure, breaking the structure-generation problem down into problems of high-level tertiary organization first, followed by local secondary structure and finally chemical detail (Fig. 4). These models exhibit the capability to implicitly model topology and syntax, choosing to allocate protein residues to different types of secondary structures during this process. With improved generation quality came the ability to outpace physics- and hallucination-based methods for rapid structure search under conditioning[20,94,95]. RFdiffusion has been used to solve diverse protein-design problems with success rates orders of magnitude higher than those of previous methods, including scaffolding motifs, generating symmetric oligomers and designing metal and protein binders[20]. This success and the success of related models exhibit the strengths of deep learning-based structure design: faster and more efficient

sampling, a high degree of automation and reproducibility and new solutions of high realism and quality.

Despite being trained on the natural distribution of protein structure, many design models sample from a distribution more aligned with design objectives that produce more globular proteins with clean topologies, syntaxes and fewer loop residues than natural proteins (Fig. 3c). It is unclear what gives rise to this latent distribution, but possibilities include regularization in neural networks and low-temperature sampling. Regularization comes in many forms, such as dropout layers, noising data augmentations and restricted model architectures, and are intended to reduce overfitting to irregularities and outliers, such as loop regions. Temperature-adjusted sampling, inspired by the role of temperature in statistical mechanics, trades off sample quality with diversity and is implemented with the temperature parameter in Chroma, the noise scale in RFdiffusion, the step scale in Protpardelle and various other strategies. Reducing the sampling temperature redistributes density from the tails of the data distribution to concentrate it at the modes, effectively exploring fewer states and focusing on high-probability ones. This is likely applied because the learned distributions fail to exactly recapitulate the natural distributions, especially in the tails, and enriching for high-quality samples is best attained by sampling at the modes. Whatever the underlying cause, it is observed empirically that these generative models sample from sharpened, centralized distributions that filter out the structural ‘noise’ present in natural proteins and yield the idealized backbones typical of de novo design, which are simpler to understand and easier to fold (Fig. 3c).

Designing sequence to specify structure and function

In the end, only a sequence is needed to describe a protein in full, but a simple string of amino acids, and the process of deducing the correct one, carries more complexity than meets the eye. When examining the sequence directly, features such as patterns of polar and hydrophobic residues as well as strategic use of glycines and prolines can be analyzed to offer a simplistic picture of protein properties, for example, its secondary-structure content or whether it can be a membrane protein. When viewed together with the structure, however, every facet of the sequence including length, pattern and amino acid identities defines an exquisite agreement with its three-dimensional structure. It may even be fair to say that, although the sequence is the ultimate expression of a protein, it is made to serve the functional purpose of the protein structure. In defining the sequence of a de novo protein, the searchable sequence space can be more extensive than that for native proteins, as structure becomes the only constraint, unbound by evolutionary requirements[59,70,96,97]. This is also true for exploration of local nearby sequences, for example, to improve the function of an enzyme[98].

Thoughtful sequence design can illuminate new insights into the interaction between protein sequence and structure. For example, the specificity of side chain packing is typically considered crucial for driving a polypeptide chain into a well-defined fold rather than a molten globule state. While investigating the influence of side chain mutations on stability, Koga et al. uncovered a counterintuitive result on hydrophobic packing specificity, namely that, in an idealized topology (in this case, a Rossmann fold), it is possible for a protein to

retain structural and thermodynamic properties despite massive mutational perturbations[99]. Specifically, despite the mutation of all buried hydrophobic residues from large to small side chains (for example, leucine or isoleucine to valine), the protein was able to not only remain folded but also retained high thermostability and an identical folded state structure.

Similar to structure design, fixed-backbone sequence design, also known as inverse folding, has also profited from deep learning and data-driven approaches. The combinatorial nature of sequence space mirrors that of protein structure, which explodes with the length of the protein and can be very difficult to search over. As with structure design, structure-prediction models provide an effective handle to grapple with this space, and the same approaches that can be used to produce protein structures from these networks can also be used to design sequences. Earlier work explored the capacity of trRosetta to define a sequence profile based on a target structure, guiding conventional methods to better conform to the global energy landscape[100]. Later, hallucination and masked inpainting methods were also found to be effective for extracting sequences from structure-prediction networks[55,56,80,101]. However, optimizing under AlphaFold2 directly with hallucination often yielded adversarial sequences, meaning sequences that AlphaFold2 predicts with high confidence but that fail to express in the wet laboratory[18,80]. The most effective sequence-design methods benefit from the strong constraint of a target structure that limits the search space: the optimal amino acid for a position is mostly determined by its local environment. This inductive bias is exploited to great effect by various types of sequence-design methods, including Gibbs and Metropolis sampling algorithms guided by physics-based or learned potentials[15,102] and masked language and autoregressive models[103,104]. These methods enable a high level of automation, generating high-quality sequences quickly with little or no manual intervention and even rescuing unfoldable sequences designed by conventional methods such as Rosetta in the case of ProteinMPNN[103].

Perhaps the most important result in protein design in the last few years is the ability to evaluate designs with the self-consistency or designability metric. Previously, computational designs were validated by ab initio structure prediction, essentially simulations of protein folding guided by an energy function, which probed the ability of a designed sequence to find the correct structure. These simulations were highly informative, offering statistical and structural insights on the impact of pathological amino acids in the sequence. However, they required large-scale computation for limited accuracy and exhibited poor correlation with experimental success. With the advent of accurate structure-prediction methods such as AlphaFold, it became possible to compare the predicted fold of a designed sequence and the original designed structure. Relatively quick computation enables predicting the folded state of a designed sequence together with a confidence metric (such as pLDDT or pAE). One might expect a sequence that is predicted to fold back to the designed structure with high confidence (‘self-consistent’ or ‘designable’) to be more consistent with the designed structure and thus potentially more likely to fold in the wet laboratory (Fig. 5)[55,60,105]. In general, these findings have substantially increased the speed and the efficiency of method development because models and designed sequences can be more faithfully evaluated in silico without requiring slower and more laborious feedback from wet laboratory validation. We also note that further work remains to improve these in silico metrics. Precision can be improved, as AlphaFold2 is susceptible to adversarial inputs, although filtering designs

through an orthogonal metric (for example, using AlphaFold2 pLDDT when sequences are hallucinated from RoseTTAFold)[55] or using structure-prediction networks that contain a language model[106] seems to mitigate this effect. The true recall is unknown because it is possible that AlphaFold2 may be 'overfit' to natural protein sequences and it is hard to estimate how often it rejects de novo sequences that would be valid, and, even within known sequence space, many natural, functional proteins cannot be predicted free of multiple sequence alignment (MSA) and would fail the self-consistency test[103].

More broadly, the structure-prediction task has proven to be very useful for protein modeling and design, with models shown to capture much more information than simply a mapping from MSAs to structures. Investigation into AlphaFold2's ability to discriminate between successful decoys in structure prediction suggests that the structure module learns a form of implicit 'energy landscape' that allows it to evaluate the plausibility of a given structure and may explain its ability to generalize to diverse tasks[107]. In addition to other applications for structure and sequence design previously discussed, AlphaFold2 has been found to be effective for predicting protein–protein interactions[105], predicting and designing cyclic peptides[108] and predicting small-molecule-binding sites[109], despite not being trained specifically for these tasks.

Finally, in a phenomenon analogous to the idealized distributions observed in structure modeling, it appears that learned sequence models also produce more 'modal' samples, perhaps due to similar effects of regularization and temperature-tuned sampling schemes. One symptom of this is that model likelihoods are typically higher on de novo-designed sequences than on natural protein sequences[102,103]. Thus, while they may have been trained to try to reproduce the natural protein distribution, a hidden de novo distribution can be extracted from within the learned natural distribution with low-temperature sampling and other strategies. To complete the cycle, this distribution of de novo proteins interacts favorably with AlphaFold2. Single-sequence (MSA-free) prediction with AlphaFold2 seemed to perform poorly with natural sequences but much better with de novo-designed protein sequences, with high self-consistency for these, enriching specifically for successful designs[103]. Why this occurs remains unknown; perhaps these de novo sequences contain more 'folding signal' (ref. 110). We observe that the highest AlphaFold2 self-consistency values and the highest rates of experimental success are achieved when sampling from sharpened (for example, idealized) structure and sequence distributions, compared to samples closer to natural distributions.

Looking beyond the central dogma

There is an old parable about a group of blind people who encounter an elephant. Each person interacts with a different part of the elephant by touch. The first handles the trunk and describes the elephant as being like a snake. The second touches the ear and decides that the elephant is some kind of fan. A third person feels the leg and declares the elephant to be the trunk of a tree. While none of them are incorrect in their observations, they all have unique perspectives that only partially touch upon the complete truth.

Like elephants, proteins also have many unique attributes and representations, although sometimes we may focus on only a single one, such as the modeling of backbone structure. The classical hierarchy of function–structure–sequence enabled progress to be made by breaking down the grand challenge of protein design into more tractable subproblems. Like the blind men in the parable, solutions to these subproblems, while effective, only consider a single facet of the many-sided nature of proteins. These approaches are more restrictive than the true nature of protein function: in real proteins, sequence influences structure through side chain interactions, function-guided evolution constrains the sequences of proteins with highly similar structures and other harder-to-observe variables such as conformational dynamics or cellular context interact with all three.

The influx of new methods coinciding with the rise of deep learning for protein design has already begun to bring new perspectives on protein modeling and design. These include approaches that directly model sequence, perhaps conditioned on or jointly with functional properties[111–114]. These protein language models have shown the ability to capture information from sequence evolution, exploring the sequence space of natural protein families and solving some protein-design tasks such as scaffolding. Their capacity to explore de novo sequence space increases substantially when they have access to structural information. Language models have been shown to learn some understanding of structure by way of coevolution, the same mechanism underpinning modern structure prediction[115–121], and, when equipped with a structure head, they have been used successfully to generate de novo proteins[119,122].

As modeling capabilities improve, it is natural to consider increasingly integrative approaches to protein design, such as all-atom modeling and co-design of protein structure and sequence. We distinguish between all-atom modeling, which is the simultaneous generation of backbone structure and side chain structure, and structure and sequence co-design, which is the simultaneous generation of structure and sequence. All-atom modeling re-emphasizes the role of side chains in protein design, which might otherwise seem like an afterthought in the central dogma but are of course critical in defining protein function. Such models enable the design of proteins with side chains considered jointly with backbone throughout the generation process, allowing for explicit modeling of chemical interactions with a target, for side chains to influence the backbone conformation and even for explicit conditioning on side chains without backbone[95]. Co-design extends this by incorporating meaningful prior distributions on sequence and could enable conditioning structure design on sequence information and vice versa or conditioning both jointly on functional information. Protein structure is degenerate, containing only a small number of unique secondary-structure types and a limited set of ways to combine these into tertiary motifs[123]. Sequence is a much richer and more expressive representation, but, as a result, its design space has been less exhaustively explored by evolution. A combination of the strengths of each approach could enable broader exploration of protein space through structure while giving fine resolution through sequence[112,113,124–126]. Repurposed structure-prediction networks immediately suggested a natural way to co-design protein structure and sequence, as the outputs include both a structure and a sequence that are self-consistent by construction[55,56,100,101], but these methods can be prone to adversarial outputs[18,80,101,103]. Generative modeling approaches were recently extended to include sequence diffusion with

structure guidance[127], all-atom structure diffusion with side chain awareness[95] and sequence and structure co-design with protein language models[119,122]. However, a true foundational model that exhibits generalization through structure and fine control over sequence, with the ability to generate and map between either one and infer causal relationships with function, a model that 'sees the whole elephant', so to speak, remains lacking.

The other 'elephant in the room' is the dynamic nature of protein structures. A strong assumption throughout de novo protein design as we have presented it here is that the folding and function of a protein is largely enthalpic in nature: structures can be designed by directly stabilizing the folded state while ignoring competing states, and function is mediated entirely through the interface with the target[110]. Of course, static structure is an incomplete model that we work with out of convenience. To date, both large-scale and fine-grained dynamic behavior and even intrinsic disorder have been achieved by design[128–132], and some consideration has been given to multiple states in sequence design[100,103,127,133], but this area remains underexplored, in part due to the paucity of ensemble and dynamics data and the difficulty of in silico evaluation. Much as how the rapid growth of the PDB has enabled advances in protein structure modeling, the development of large, standardized conformation datasets should unlock further progress in this area[134]. Reincorporating physics-based inductive biases into existing deep learning models may also allow for generalizing to conformations while leaning only on static structure data.

Examining specific applications of protein-design methods (examples shown in Fig. 6), two areas of interest are the design of antibodies and enzymes. Both applications present unique challenges for current design methods. Antibodies are ubiquitous as a therapeutic modality, and the ability to design them relatively quickly and cheaply compared to animal immunization would be of substantial impact[135]. In comparison to de novo protein binders, antibodies typically affect binding through loop-rich complementarity-determining regions, which are difficult to model compared to helices and sheets for both design and evaluation methods such as AlphaFold2 (refs. 1,136). It is of similar importance to control the developability properties of antibody sequences to avoid oligomerization and other behaviors that interfere with their ability to function as therapeutics[137].

Enzymes are frequently sought after in applications in which catalysts that can unlock new chemical transformations or function efficiently in mild conditions would contribute to sustainability, new materials and synthesis pathways. These molecules also present a difficult challenge for structure-based de novo design, as the sub-angstrom scale of the physical process, bond breaking and forming, requires a degree of accuracy that is not always attainable in structural datasets or with protein-design methods. Recent advances include scaffold recombination and hallucination to generate diverse solutions for placing catalytic motifs[138,139]. In some cases, manual intervention with information from evolution has been required[98]. Accurate modeling of multiple conformational states would likely be beneficial to both antibody and enzyme design, for example, the effects of complementarity-determining region flexibility on antibody binding affinity and modeling changes in active site geometry given proximal or distal mutations and their effect on catalytic activity. Recent methods for conformational sampling such as EigenFold, Distributional Graphormer,

PepFlow and MSA subsampling and clustering with AlphaFold can build toward methods in this direction[140–144].

Finally, a remaining broad challenge is the complexity of affecting phenotypes in cells and organisms with protein design. While our capacity to design functional proteins has increased rapidly in recent years, solving the molecular recognition component remains only a small part of this challenge. For example, affecting chromatin organization or gene expression may involve deeper insights than simply nucleic acid or histone binding. To direct the behavior of cells, receptor binding is often only the first challenge to be solved. Of direct interest and impact to society is the potential of designed proteins to function as therapeutics. For de novo proteins to be useful as drugs, they will need to exhibit serum stability, minimal immunogenicity and other developability traits.

Conclusion

New end-to-end pipelines reduce the labor needed to design proteins, democratizing protein design, lowering the barrier to entry and enriching the space of experiments that can be tried. We expect capabilities and performance to continue to grow as modeling techniques continue to improve and the field generates more data, making headway toward functional de novo design (Fig. 6). At the same time, it is hard to understate the need for continued biophysical learning in this new age of automated protein design. Nature remains full of complex biological systems that we are only beginning to understand, and, in the absence of perfect data, the only way to generalize to all cases is with a causal model. As is often true in science and in protein design, the first forays into uncharted directions of research are opened by careful observation and reasoning from first principles, and we expect it to remain so moving forward.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank S. Ovchinnikov for feedback on the manuscript. For readers interested in more depth on physics-based modeling approaches, such as Rosetta, we recommend other reviews[2,3,17]. We also recommend two recent reviews with related perspectives, focusing more on the details and impact of machine learning on protein engineering and design, especially on direct sequence modeling[73,145]. A.E.C. is supported by the NSF GRFP and the Merck SEEDS Program. T.L. is supported by a Stanford Graduate Fellowship. P.-S.H. is supported by the NIH (R01GM147893), the American Cancer Society (ACS 134055-IRG-218), the BASF CARA project and the Discovery Innovation Fund.

References

1. Chothia C Principles that determine the structure of proteins. Annu. Rev. Biochem. 53, 537–572 (1984). [PubMed: 6383199]
2. Korendovych IV & DeGrado WF De novo protein design, a retrospective. Q. Rev. Biophys. 53, e3 (2020). [PubMed: 32041676]
3. Huang P-S, Boyken SE & Baker D The coming of age of de novo protein design. Nature 537, 320–327 (2016). [PubMed: 27629638]

4. Baker D What has de novo protein design taught us about protein folding and biophysics? Protein Sci. 28, 678–683 (2019). [PubMed: 30746840]
5. DeGrado WF, Summa CM, Pavone V, Nastri F & Lombardi A De novo design and structural characterization of proteins and metalloproteins. Annu. Rev. Biochem. 68, 779–819 (1999). [PubMed: 10872466]
6. Regan L & DeGrado WF Characterization of a helical protein designed from first principles. Science 241, 976–978 (1988). [PubMed: 3043666]
7. Harbury PB, Plecs JJ, Tidor B, Alber T & Kim PS High-resolution protein design with backbone freedom. Science 282, 1462–1467 (1998). [PubMed: 9822371]
8. Dahiyat BI & Mayo SL De novo protein design: fully automated sequence selection. Science 278, 82–87 (1997). [PubMed: 9311930]
9. Dahiyat BI & Mayo SL Protein design automation. Protein Sci. 5, 895–903 (1996). [PubMed: 8732761]
10. Walsh STR, Cheng H, Bryson JW, Roder H & DeGrado WF Solution structure and dynamics of a de novo designed three-helix bundle protein. Proc. Natl Acad. Sci. USA 96, 5486–5491 (1999). [PubMed: 10318910]
11. Levinthal C Are there pathways for protein folding? J. Chim. Phys. 65, 44–45 (1968).
12. Maynard Smith J Natural selection and the concept of a protein space. Nature 225, 563–564 (1970). [PubMed: 5411867]
13. Kuhlman B et al. Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368 (2003). [PubMed: 14631033]
14. Gront D, Kulp DW, Vernon RM, Strauss CEM & Baker D Generalized fragment picking in Rosetta: design, protocols and applications. PLoS ONE 6, e23294 (2011). [PubMed: 21887241]
15. Alford RF et al. The Rosetta all-atom energy function for macromolecular modeling and design. J. Chem. Theory Comput. 13, 3031–3048 (2017). [PubMed: 28430426]
16. Shapovalov MV & Dunbrack RL A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19, 844–858 (2011). [PubMed: 21645855]
17. Leman JK et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat. Methods 17, 665–680 (2020). [PubMed: 32483333]
18. Wicky BIM et al. Hallucinating symmetric protein assemblies. Science 378, 56–61 (2022). [PubMed: 36108048]
19. Vorobieva AA et al. De novo design of transmembrane β barrels. Science 371, eabc8182 (2021). [PubMed: 33602829]
20. Watson JL et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023). [PubMed: 37433327]
21. Krishna R et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Preprint at bioRxiv 10.1101/2023.10.09.561603 (2023).
22. Sheffler W et al. Fast and versatile sequence-independent protein docking for nanomaterials design using RPXDock. PLoS Comput. Biol. 19, e1010680 (2023). [PubMed: 37216343]
23. Eguchi RR, Choe CA & Huang P-S Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. PLoS Comput. Biol. 18, e1010271 (2022). [PubMed: 35759518]
24. Lin Y & Alquraishi M Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. In Proceedings of the 40th International Conference on Machine Learning (eds. Krause A et al.) Vol. 202, 20978–21002 (PMLR, 2023); https://proceedings.mlr.press/v202/lin23a.html
25. Wu KE et al. Protein structure generation via folding diffusion. Preprint at arXiv 10.48550/arXiv.2209.15611 (2022).
26. Yim J et al. SE(3) diffusion model with application to protein backbone generation. In Proceedings of the 40th International Conference on Machine Learning (eds. Krause A et al.) Vol. 202, 40001–40039 (PMLR, 2023); https://proceedings.mlr.press/v202/yim23a.html
27. Bose JA et al. SE(3)-stochastic flow matching for protein backbone generation. Preprint at arXiv 10.48550/arXiv.2310.02391 (2024).

28. Yim J et al. Fast protein backbone generation with SE(3) flow matching. Preprint at arXiv 10.48550/arXiv.2310.05297 (2023).
29. Fu C et al. A latent diffusion model for protein structure generation. Preprint at arXiv 10.48550/arXiv.2305.04120 (2023).
30. Liu Y, Chen L & Liu H Diffusion in a quantized vector space generates non-idealized protein structures and predicts conformational distributions. Preprint at arXiv 10.1101/2023.11.18.567666 (2023).
31. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A & Kim PM Fast and flexible protein design using deep graph neural networks. Cell Syst. 11, 402–411 (2020). [PubMed: 32971019]
32. Ingraham J, Garg V, Barzilay R & Jaakkola T Generative models for graph-based protein design. In Advances in Neural Information Processing Systems (eds. Wallach H et al.) Vol. 32. (Curran Associates, 2019); https://proceedings.neurips.cc/paper_files/paper/2019/file/f3a4ff4839c56a5f460c88cce3666a2b-Paper.pdf
33. Gao Z et al. PiFold: toward effective and efficient protein inverse folding. Preprint at arXiv 10.48550/arXiv.2209.12643 (2022).
34. Yi K et al. Graph denoising diffusion for inverse protein folding. Preprint at arXiv 10.48550/arXiv.2306.16819 (2023).
35. Hsu C et al. Learning inverse folding from millions of predicted structures. In Proceedings of the 39th International Conference on Machine Learning (eds. Chaudhuri K et al.) Vol. 162, 8946–8970 (PMLR, 2022); https://proceedings.mlr.press/v162/hsu22a.html
36. Xiong P et al. Increasing the efficiency and accuracy of the ABACUS protein sequence design method. Bioinformatics 36, 136–144 (2020). [PubMed: 31240299]
37. Liu Y et al. Rotamer-free protein sequence design based on deep learning and self-consistency. Nat. Comput. Sci. 2, 451–462 (2022). [PubMed: 38177863]
38. Heinzinger M et al. ProstT5: bilingual language model for protein sequence and structure. Preprint at bioRxiv 10.1101/2023.07.23.550085 (2023).
39. Su J et al. SaProt: protein language modeling with structure-aware vocabulary. Preprint at bioRxiv 10.1101/2023.10.01.560349 (2023).
40. Gruver N et al. Protein design with guided discrete diffusion. Preprint at arXiv 10.48550/arXiv.2305.20009 (2023).
41. Repecka D et al. Expanding functional protein sequence spaces using generative adversarial networks. Nat. Mach. Intell. 3, 324–333 (2021).
42. Greener JG, Moffat L & Jones DT Design of metalloproteins and novel protein folds using variational autoencoders. Sci. Rep. 8, 16189 (2018). [PubMed: 30385875]
43. Jin W, Wohlwend J, Barzilay R & Jaakkola T Iterative refinement graph neural network for antibody sequence–structure co-design. Preprint at arXiv 10.48550/arXiv.2110.04624 (2022).
44. Martinkus K et al. AbDiffuser: full-atom generation of in-vitro functioning antibodies. Preprint at arXiv 10.48550/arXiv.2308.05027 (NeurIPS, 2023).
45. Luo S Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In Advances in Neural Information Processing Systems (eds. Koyejo S et al.) Vol. 35, 9754–9767 (Curran Associates, Inc., 2022); https://proceedings.neurips.cc/paper_files/paper/2022/file/3fa7d76a0dc1179f1e98d1bc62403756-Paper-Conference.pdf
46. Davison J Zero-shot learning in modern NLP. Joe Davison Blog joeddav.github.io/blog/2020/05/29/ZSL.html (2020).
47. Chen RTQ, Rubanova Y, Bettencourt J & Duvenaud DK Neural ordinary differential equations. In Advances in Neural Information Processing Systems (eds. Bengio S et al.) Vol. 31 (Curran Associates, 2018); https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf
48. Lipman Y, Chen RT, Ben-Hamu H, Nickel M & Le M Flow matching for generative modeling. Preprint at arXiv 10.48550/arXiv.2210.02747 (2023).
49. Liu X, Gong C & Liu Q Flow straight and fast: learning to generate and transfer data with rectified flow. Preprint at arXiv 10.48550/arXiv.2209.03003 (2022).

50. Albergo MS, Boffi NM & Vanden-Eijnden E Stochastic interpolants: a unifying framework for flows and diffusions. Preprint at arXiv 10.48550/arXiv.2303.08797 (2023).
51. Somnath VR et al. Aligned diffusion Schrödinger bridges. Preprint at arXiv 10.48550/arXiv.2302.11419 (2023).
52. Conte LL, Chothia C & Janin J The atomic structure of protein–protein recognition sites. J. Mol. Biol. 285, 2177–2198 (1999). [PubMed: 9925793]
53. Woolfson DN A brief history of de novo protein design: minimal, rational, and computational. J. Mol. Biol. 433, 167160 (2021). [PubMed: 34298061]
54. Sesterhenn F et al. De novo protein design enables the precise induction of RSV-neutralizing antibodies. Science 368, eaay5051 (2020). [PubMed: 32409444]
55. Wang J et al. Scaffolding protein functional sites using deep learning. Science 377, 387–394 (2022). [PubMed: 35862514]
56. Anishchenko I et al. De novo protein design by deep network hallucination. Nature 600, 547–552 (2021). [PubMed: 34853475]
57. Scott AJ et al. Constructing ion channels from water-soluble α-helical barrels. Nat. Chem. 13, 643–650 (2021). [PubMed: 33972753]
58. Mahendran KR et al. A monodisperse transmembrane α-helical peptide barrel. Nat. Chem. 9, 411–419 (2017). [PubMed: 28430192]
59. Dou J et al. De novo design of a fluorescence-activating β-barrel. Nature 561, 485–491 (2018). [PubMed: 30209393]
60. Cao L et al. Design of protein-binding proteins from the target structure alone. Nature 605, 551–560 (2022). [PubMed: 35332283]
61. Polizzi NF & DeGrado WF A defined structural unit enables de novo design of small-molecule-binding proteins. Science 369, 1227–1233 (2020). [PubMed: 32883865]
62. Eguchi RR et al. Deep generative design of epitope-specific binding proteins by latent conformation optimization. Preprint at bioRxiv 10.1101/2022.12.22.521698 (2022).
63. Glasscock CJ et al. Computational design of sequence-specific DNA-binding proteins. Preprint at bioRxiv 10.1101/2023.09.20.558720 (2023).
64. Gainza P et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat. Methods 17, 184–192 (2020). [PubMed: 31819266]
65. Gainza P et al. De novo design of protein interactions with learned surface fingerprints. Nature 617, 176–184 (2023). [PubMed: 37100904]
66. Torres SV et al. De novo design of high-affinity binders of bioactive helical peptides. Nature 10.1038/s41586-023-06953-1 (2023).
67. Koga N et al. Principles for designing ideal protein structures. Nature 491, 222–227 (2012). [PubMed: 23135467]
68. Chu AE, Fernandez D, Liu J, Eguchi RR & Huang P-S De novo design of a highly stable ovoid TIM barrel: unlocking pocket shape towards functional design. Biodes. Res. 2022, 9842315 (2022). [PubMed: 37850141]
69. Huang P-S et al. RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS ONE 6, e24109 (2011). [PubMed: 21909381]
70. Marcos E et al. De novo design of a non-local β-sheet protein with high stability and accuracy. Nat. Struct. Mol. Biol. 25, 1028–1034 (2018). [PubMed: 30374087]
71. Huang P-S et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat. Chem. Biol. 12, 29–34 (2016). [PubMed: 26595462]
72. Jiang L et al. De novo computational design of retro-aldol enzymes. Science 319, 1387–1391 (2008). [PubMed: 18323453]
73. Winnifrith A, Outeiral C & Hie B Generative artificial intelligence for de novo protein design. Preprint at arXiv 10.48550/arXiv.2310.09685 (2023).
74. Huang B et al. A backbone-centred energy function of neural networks for protein design. Nature 602, 523–528 (2022). [PubMed: 35140398]
75. Senior AW et al. Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710 (2020). [PubMed: 31942072]

76. Yang J et al. Improved protein structure prediction using predicted interresidue orientations. Proc. Natl Acad. Sci. USA 117, 1496–1503 (2020). [PubMed: 31896580]
77. Jumper J et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). [PubMed: 34265844]
78. Baek M et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876 (2021). [PubMed: 34282049]
79. Baek M et al. Efficient and accurate prediction of protein structure using RoseTTAFold2. Preprint at bioRxiv 10.1101/2023.05.24.542179 (2023).
80. Frank C et al. Efficient and scalable de novo protein design using a relaxed sequence space. Preprint at bioRxiv 10.1101/2023.02.24.529906 (2023).
81. Tischer D et al. Design of proteins presenting discontinuous functional sites using deep learning. Preprint at bioRxiv 10.1101/2020.11.29.402743 (2020).
82. Goodfellow I et al. Generative adversarial networks. Commun. ACM 63, 139–144 (2020).
83. Radford A et al. Language models are unsupervised multitask learners. OpenAI Blog 1, 9 (2019).
84. Radford A, Metz L & Chintala S Unsupervised representation learning with deep convolutional generative adversarial networks. Preprint at arXiv 10.48550/arXiv.1511.06434 (2015).
85. Kingma DP & Welling M Auto-encoding variational Bayes. Preprint at arXiv 10.48550/arXiv.1312.6114 (2013).
86. Karras T, Laine S & Aila T A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 4401–4410 (IEEE, 2018).
87. Anand N & Huang P Generative modeling for protein structures. In Advances in Neural Information Processing Systems (eds. Bengio S et al.) Vol. 31 (Curran Associates, 2018); https://proceedings.neurips.cc/paper_files/paper/2018/file/afa299a4d1d8c52e75dd8a24c3ce534f-Paper.pdf
88. Ho J, Jain A & Abbeel P Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (eds. Larochelle H et al.) Vol. 33, 6840–6851 (Curran Associates, 2020); https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
89. Song Y et al. Score-based generative modeling through stochastic differential equations. Preprint at arXiv 10.48550/arXiv.2011.13456 (2021).
90. Dhariwal P & Nichol A Diffusion models beat GANs on image synthesis. In Advances in Neural Information Processing Systems (eds. Ranzato M et al.) Vol. 34, 8780–8794 (Curran Associates, 2021); https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf
91. Anand N & Achim T Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. Preprint at arXiv 10.48550/arXiv.2205.15019 (2022).
92. Li CT & Farnia F Mode-seeking divergences: theory and applications to GANs. In Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (eds. Ruiz F, Dy J & van de Meent J-W) Vol. 206, 8321–8350 (PMLR, 2023); https://proceedings.mlr.press/v206/ting-li23a.html
93. Lee JS, Kim J & Kim PM Score-based generative modeling for de novo protein design. Nat. Comput. Sci. 3, 382–392 (2023). [PubMed: 38177840]
94. Ingraham JB et al. Illuminating protein space with a programmable generative model. Nature 623, 1070–1078 (2023). [PubMed: 37968394]
95. Chu AE, Cheng L, Nesr GE, Xu M & Huang P-S An all-atom protein generative model. Preprint at bioRxiv 10.1101/2023.05.24.542194 (2023).
96. Basanta B et al. An enumerative algorithm for de novo design of proteins with diverse pocket structures. Proc. Natl Acad. Sci. USA 117, 22135–22145 (2020). [PubMed: 32839327]
97. Mravic M et al. Packing of apolar side chains enables accurate design of highly stable membrane proteins. Science 363, 1418–1423 (2019). [PubMed: 30923216]
98. Sumida KH et al. Improving protein expression, stability, and function with ProteinMPNN. Preprint at bioRxiv 10.1101/2023.10.03.560713 (2023).

99. Koga R et al. Robust folding of a de novo designed ideal protein even with most of the core mutated to valine. Proc. Natl Acad. Sci. USA 117, 31149–31156 (2020). [PubMed: 33229587]
100. Norn C et al. Protein sequence design by conformational landscape optimization. Proc. Natl Acad. Sci. USA 118, e2017228118 (2021). [PubMed: 33712545]
101. Goverde CA, Wolf B, Khakzad H, Rosset S & Correia BE De novo protein design by inversion of the AlphaFold structure prediction network. Protein Sci. 32, e4653 (2023). [PubMed: 37165539]
102. Anand N et al. Protein sequence design with a learned potential. Nat. Commun. 13, 746 (2022). [PubMed: 35136054]
103. Dauparas J et al. Robust deep learning-based protein sequence design using ProteinMPNN. Science 378, 49–56 (2022). [PubMed: 36108050]
104. Yang KK, Zanichelli N & Yeh H Masked inverse folding with sequence transfer for protein representation learning. Protein Eng. Des. Sel. 36, gzad015 (2023). [PubMed: 37883472]
105. Evans R et al. Protein complex prediction with AlphaFold-Multimer. Preprint at bioRxiv 10.1101/2021.10.04.463034 (2022).
106. Jeliazkov JR, Alamo Ddel & Karpiak JD ESMFold hallucinates native-like protein sequences. In NeurIPS Workshop on Machine Learning in Structural Biology. Preprint at bioRxiv 10.1101/2023.05.23.541774 (2023).
107. Rettie SA et al. Cyclic peptide structure prediction and design using AlphaFold. Preprint at bioRxiv 10.1101/2023.02.25.529956 (2023).
108. Roney JP & Ovchinnikov S State-of-the-art estimation of protein model accuracy using AlphaFold. Phys. Rev. Lett. 129, 238101 (2022). [PubMed: 36563190]
109. Gazizov A, Lian A, Goverde C, Ovchinnikov S & Polizzi NF AF2BIND: predicting ligand-binding sites using the pair representation of AlphaFold2. Preprint at bioRxiv 10.1101/2023.10.15.562410 (2023).
110. Fleishman SJ & Baker D Role of the biomolecular energy gap in protein design, structure, and evolution. Cell 149, 262–273 (2012). [PubMed: 22500796]
111. Madani A et al. Large language models generate functional protein sequences across diverse families. Nat. Biotechnol. 41, 1099–1106 (2023). [PubMed: 36702895]
112. Nijkamp E, Ruffolo J, Weinstein EN, Naik N & Madani A ProGen2: exploring the boundaries of protein language models. Cell Syst. 14, 968–978 (2023). [PubMed: 37909046]
113. Ferruz N, Schmidt S & Höcker B ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun. 13, 4348 (2022). [PubMed: 35896542]
114. Alamdari S et al. Protein generation with evolutionary diffusion: sequence is all you need. Preprint at bioRxiv 10.1101/2023.09.11.556673 (2023).
115. Kamisetty H, Ovchinnikov S & Baker D Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc. Natl Acad. Sci. USA 110, 15674–15679 (2013). [PubMed: 24009338]
116. Rao R et al. Evaluating protein transfer learning with TAPE. In Advances in Neural Information Processing Systems (eds. Wallach H et al.) Vol. 32 (Curran Associates, 2019); https://proceedings.neurips.cc/paper_files/paper/2019/file/37f65c068b7723cd7809ee2d31d7861c-Paper.pdf
117. Vig J et al. BERTology meets biology: interpreting attention in protein language models. Preprint at arXiv 10.48550/arXiv.2006.15222 (2021).
118. Bepler T & Berger B Learning the protein language: evolution, structure, and function. Cell Syst. 12, 654–669 (2021). [PubMed: 34139171]
119. Verkuil R et al. Language models generalize beyond natural proteins. Preprint at bioRxiv 10.1101/2022.12.21.521521 (2022).
120. Wu R et al. High-resolution de novo structure prediction from primary sequence. Preprint at bioRxiv 10.1101/2022.07.21.500999 (2022).
121. Lin Z et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379, 1123–1130 (2023). [PubMed: 36927031]
122. Hie B et al. A high-level programming language for generative protein design. Preprint at bioRxiv 10.1101/2022.12.21.521526 (2022).

123. Mackenzie CO, Zhou J & Grigoryan G Tertiary alphabet for the observable protein structural universe. Proc. Natl Acad. Sci. USA 113, E7438–E7447 (2016). [PubMed: 27810958]
124. Riesselman AJ, Ingraham JB & Marks DS Deep generative models of genetic variation capture the effects of mutations. Nat. Methods 15, 816–822 (2018). [PubMed: 30250057]
125. Shin JE et al. Protein design and variant prediction using autoregressive generative models. Nat. Commun. 12, 2403 (2021). [PubMed: 33893299]
126. Brookes D, Park H & Listgarten J Conditioning by adaptive sampling for robust design. In Proceedings of the 36th International Conference on Machine Learning (eds. Chaudhuri K & Salakhutdinov R) Vol. 97, 773–782 (PMLR, 2019); https://proceedings.mlr.press/v97/brookes19a.html
127. Lisanza SL et al. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion. Preprint at bioRxiv 10.1101/2023.05.08.539766 (2023).
128. Langan RA et al. De novo design of bioactive protein switches. Nature 572, 205–210 (2019). [PubMed: 31341284]
129. Praetorius F et al. Design of stimulus-responsive two-state hinge proteins. Science 381, 754–760 (2023). [PubMed: 37590357]
130. Wei KY et al. Computational design of closely related proteins that adopt two well-defined but structurally divergent folds. Proc. Natl Acad. Sci. USA 117, 7208–7215 (2020). [PubMed: 32188784]
131. St-Jacques AD et al. Computational remodeling of an enzyme conformational landscape for altered substrate selectivity. Nat. Commun. 14, 6058 (2023). [PubMed: 37770431]
132. Pesce F et al. Design of intrinsically disordered protein variants with diverse structural properties. Preprint at bioRxiv 10.1101/2023.10.22.563461 (2023).
133. Leaver-Fay A, Jacak R, Stranges PB & Kuhlman B A generic program for multistate protein design. PLoS ONE 6, e20937 (2011). [PubMed: 21754981]
134. Wankowicz SA et al. Uncovering protein ensembles: automated multiconformer model building for X-ray crystallography and cryo-EM. Preprint at bioRxiv 10.1101/2023.06.28.546963 (2023).
135. Kim J, McFee M, Fang Q, Abdin O & Kim PM Computational and artificial intelligence-based methods for antibody development. Trends Pharmacol. Sci. 44, 175–189 (2023). [PubMed: 36669976]
136. North B, Lehmann A & Dunbrack RL A new clustering of antibody CDR loop conformations. J. Mol. Biol. 406, 228–256 (2011). [PubMed: 21035459]
137. Raybould MI et al. Five computational developability guidelines for therapeutic antibody profiling. Proc. Natl Acad. Sci. USA 116, 4025–4030 (2019). [PubMed: 30765520]
138. Lipsh-Sokolik R et al. Combinatorial assembly and design of enzymes. Science 379, 195–201 (2023). [PubMed: 36634164]
139. Yeh AHW et al. De novo design of luciferases using deep learning. Nature 614, 774–780 (2023). [PubMed: 36813896]
140. Jing B et al. EigenFold: generative protein structure prediction with diffusion models. Preprint at arXiv 10.48550/arXiv.2304.02198 (2023).
141. Zheng S et al. Towards predicting equilibrium distributions for molecular systems with deep learning. Preprint at arXiv 10.48550/arXiv.2306.05445 (2023).
142. Abdin O & Kim PM PepFlow: direct conformational sampling from peptide energy landscapes through hypernetwork-conditioned diffusion. Preprint at bioRxiv 10.1101/2023.06.25.546443 (2023).
143. Wallner B AFsample: improving multimer prediction with AlphaFold using massive sampling. Bioinformatics 39, btad573 (2023). [PubMed: 37713472]
144. Wayment-Steele HK et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 10.1038/s41586-023-06832-9 (2023).
145. Khakzad H et al. A new age in protein design empowered by deep learning. Cell Syst. 14, 925–939 (2023). [PubMed: 37972559]
146. Minami S et al. Exploration of novel αβ-protein folds through de novo design. Nat. Struct. Mol. Biol. 30, 1132–1140 (2023). [PubMed: 37400653]

147. Bonet J et al. Rosetta FunFolDes — a general framework for the computational design of functional proteins. PLoS Comput. Biol. 14, e1006623 (2018). [PubMed: 30452434]

148. Dieleman S Diffusion Models are Autoencoders https://sander.ai/2022/01/31/diffusion.html (2022).

149. Boyken SE et al. De novo design of tunable, pH-driven conformational changes. Science 364, 658–664 (2019). [PubMed: 31097662]

150. Bethel NP et al. Precisely patterned nanofibres made from extendable protein multiplexes. Nat. Chem. 15, 1664–1671 (2023). [PubMed: 37667012]

151. Kurihara K et al. Crystal structure and activity of a de novo enzyme, ferric enterobactin esterase Syn-F4. Proc. Natl Acad. Sci. USA 120, e2218281120 (2023). [PubMed: 37695900]

152. Naudin EA et al. Acyl transfer catalytic activity in de novo designed protein with N-terminus of α-helix as oxyanion-binding site. J. Am. Chem. Soc. 143, 3330–3339 (2021). [PubMed: 33635059]

153. Mulligan VK et al. Computational design of mixed chirality peptide macrocycles with internal symmetry. Protein Sci. 29, 2433–2445 (2020). [PubMed: 33058266]

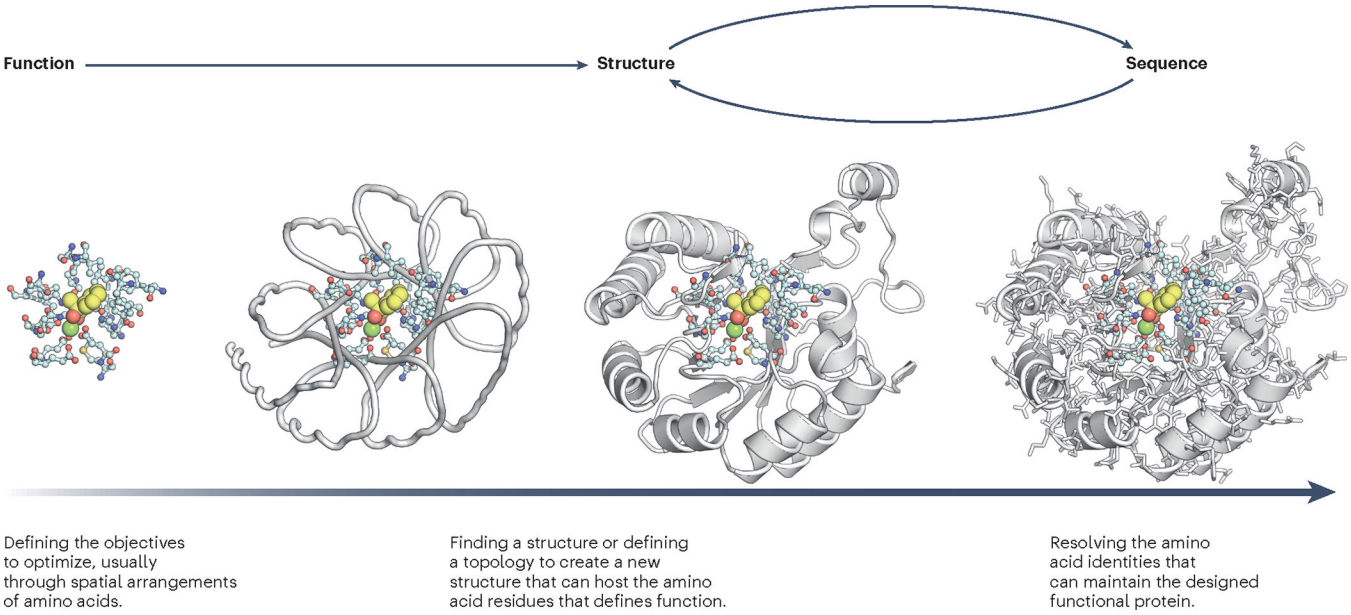

Fig. 1 |. The central dogma of de novo protein design.

Protein design generally follows this workflow: specify a desired function, design a structure that can structurally host this function, and find a sequence that folds into this structure. This central dogma underlies nearly all de novo protein-design efforts.

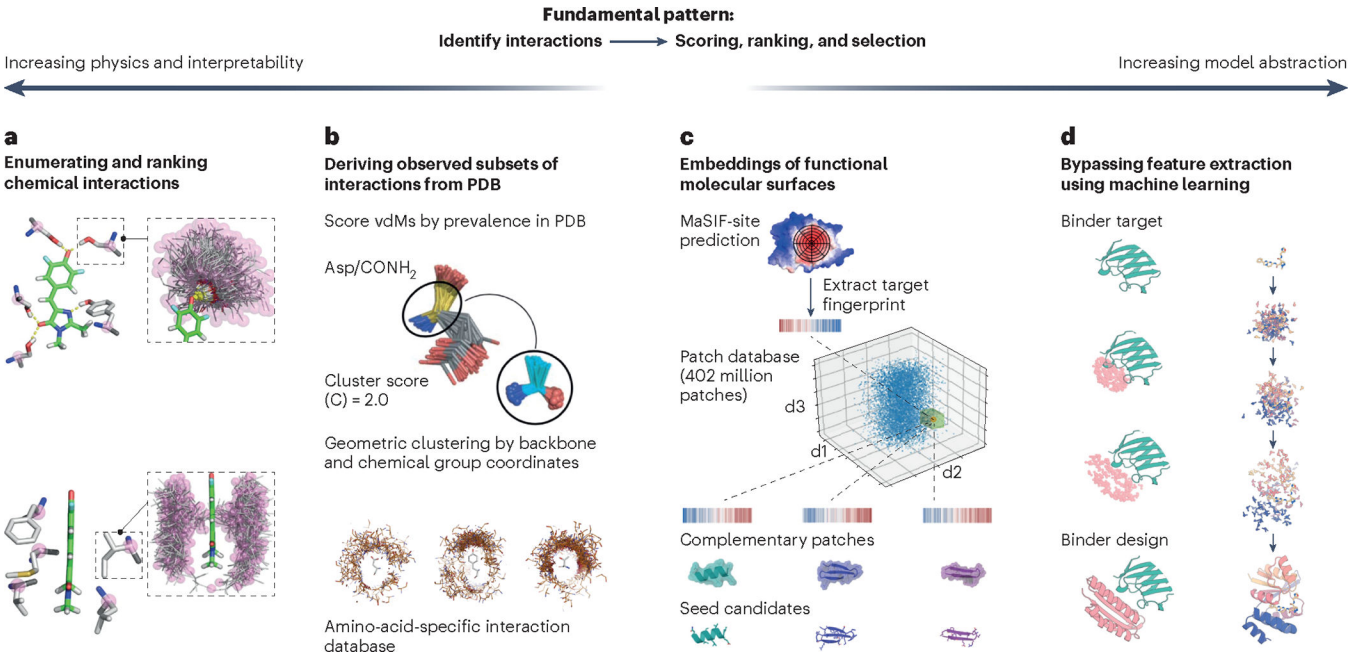

Fig. 2 |. Defining functional motifs in protein design.

A spectrum of approaches for extracting functional motifs from physics-inspired or data-driven models. **a**, Rotamer interacting fields (RIFgen and RIFdock) enumerate the space of favorable chemical interactions, scored using an explicit energy function, and use inverse rotamers to successively generate each torsion angle of the side chain up to the backbone. **b**, Extracting observed chemical interactions from the PDB, each termed a van der Mer (vdM), and scoring them with a combination of energies and statistical enrichment. COMBS applies this approach to identify backbone–ligand interacting chemical groups, and Sculptor identifies favorable protein–protein binders. **c**, Machine learning models such as MaSIF can be trained on protein and chemical data to learn high-level embeddings of functional surfaces, which can then be used to score structural elements on their complementarity. d1–d3, dimensions 1–3. **d**, With the appropriate pretraining task, generative models can directly learn the nature of protein–protein and protein–ligand interactions and sample according to these patterns. Illustrations adapted from refs. 20,59,60,64, Springer Nature; adapted with permission from ref. 61, AAAS; adapted with permission from ref. 21, CC-BY-ND 4.0; and adapted from ref. 62.

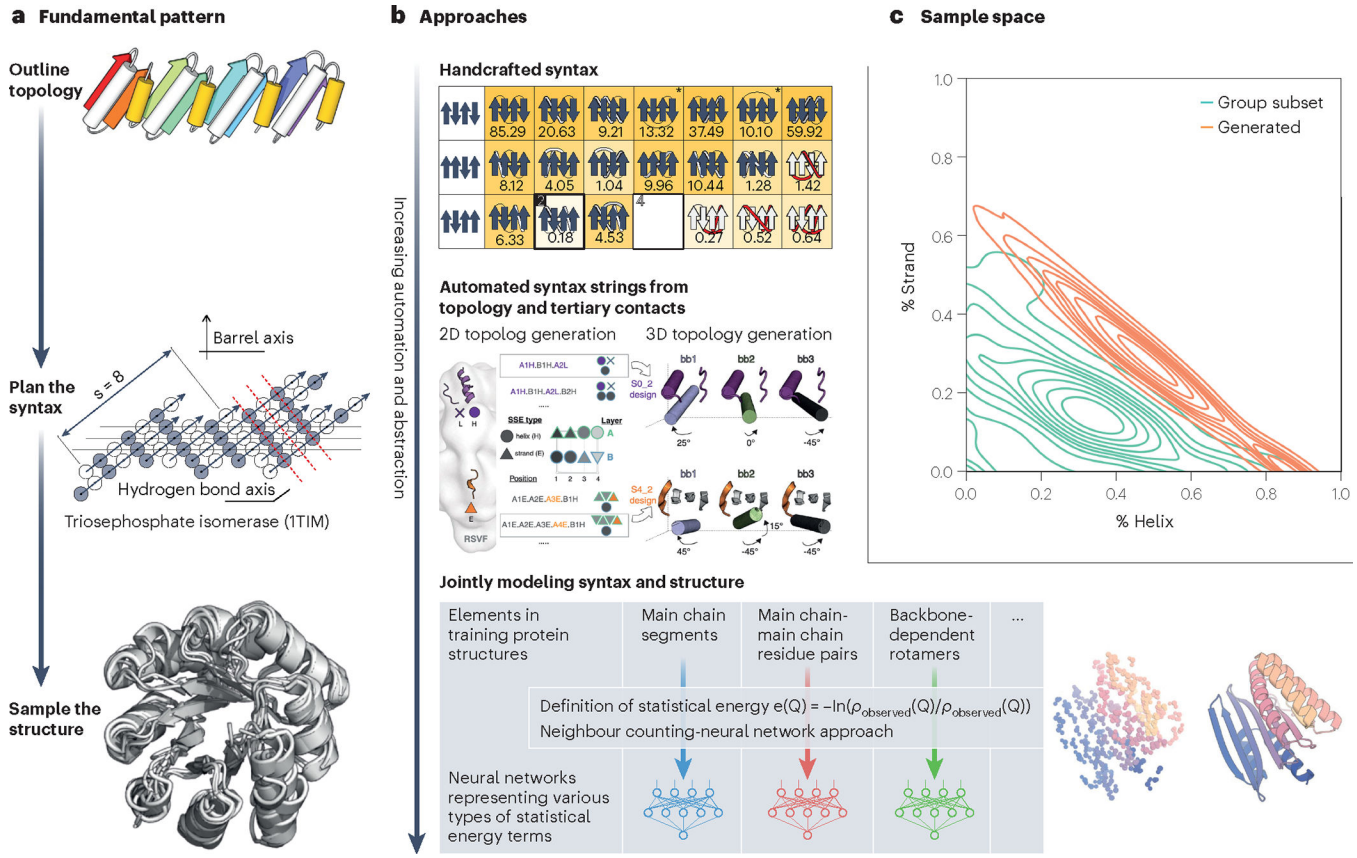

Fig. 3 |. Controlling protein structure to scaffold functional elements.

a, The basic approach of de novo structure design. First, a topology, or arrangement of secondary-structure elements, is defined. Next, the lengths of these elements, which heavily influence the tertiary structure, must be defined (the syntax). Finally, a structure can be sampled conditioned on this syntax secondary-structure string using parametric equations, Monte Carlo fragment assembly or generative models. Abbreviation: *s*, shear number. **b**, The tradeoff between design and automation. With fully handcrafted syntaxes, the protein designer has full control over the design process and can specify unique and new topologies and structures, but this approach requires substantial manual curation and inspection. More automated methods that allow more complex parametric models to sample the topology, syntax and structure with minimal human intervention may be more accessible, reproducible and efficient but afford less control over the design process and are less interpretable. In between are methods that integrate some user-specified information and build up the remaining structure based on physics and statistics. Abbreviations: 2D, two dimensional; 3D, three dimensional; SSE, secondary structure element; ρ, probability density; Q, set of structural variables. **c**, Generated structures from RFdiffusion (orange contours) show enrichment of secondary-structure content compared to those from the PDB (green), which have more regions lacking in secondary structure. Illustrations adapted from refs. 20,74,146, Springer Nature, and adapted from ref. 147, CC-BY 4.0.

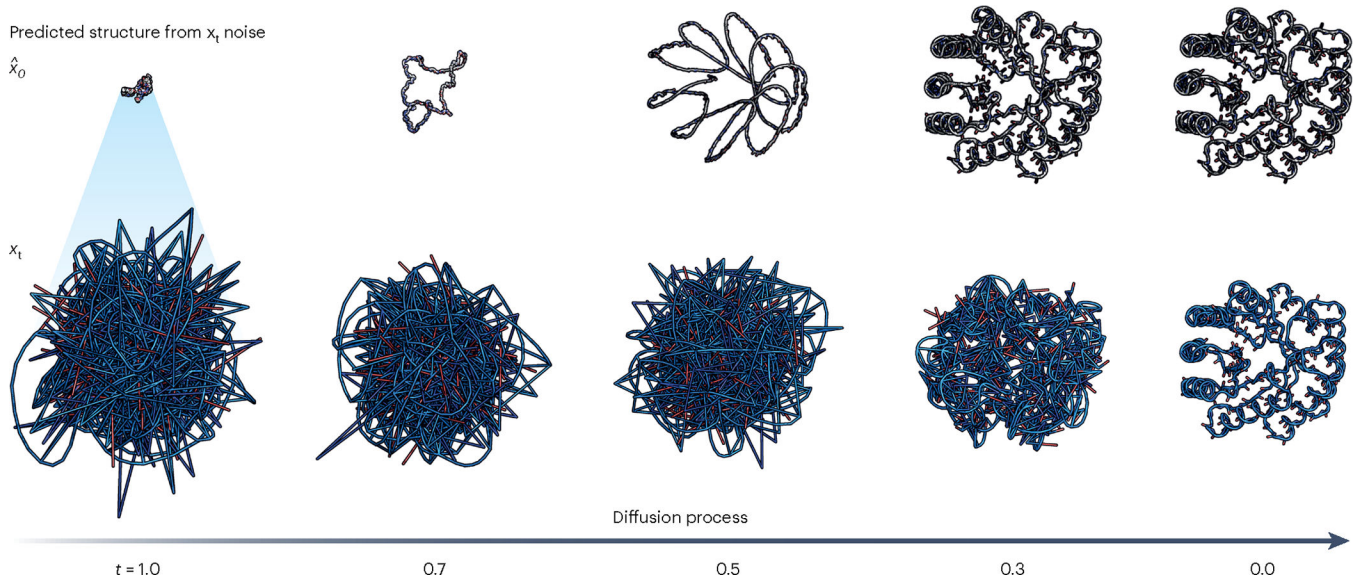

Fig. 4 |. Hierarchical nature of diffusion models.

An example diffusion sampling trajectory from backbone-only Protpardelle, which diffuses with Gaussian noise directly on the Cartesian coordinates of the protein backbone atoms[95]. The parameter t represents the proportion of noise that has been added; therefore, $t = 0$ indicates no noise, $t = 1$ indicates maximum noise, x_t represents the inputs to the denoising network at timestep t and $\hat{x}_0 = f_\theta(x_t)$ represents the network predictions of clean or denoised data, where θ denotes the parameters of the network f, that is, what the data should look like at $t = 0$, given the inputs x_t. Note that the absolute value of t may not correspond exactly to the noise variance ($\sigma = f(t)$ in general); for example, for this particular diffusion process, the standard deviation of noise added $\sigma = 800$ when $t = 1$; therefore, we downscale the coordinates to fit in the context of this figure. While to the human eye very little structural detail is apparent in the sampling trajectory (x_t) until late in the denoising process, we can see that the model extracts signal (x_0) in a way that proceeds from ‘low-frequency’ information (that is, tertiary organization, which involves many atoms) to ‘medium-frequency’ features (for example, secondary structure, which involves fewer atoms) and eventually ‘high-frequency’ details (such as bond lengths and angles, which involves only a few atoms). For further discussion, see section 5 of ref. 148.

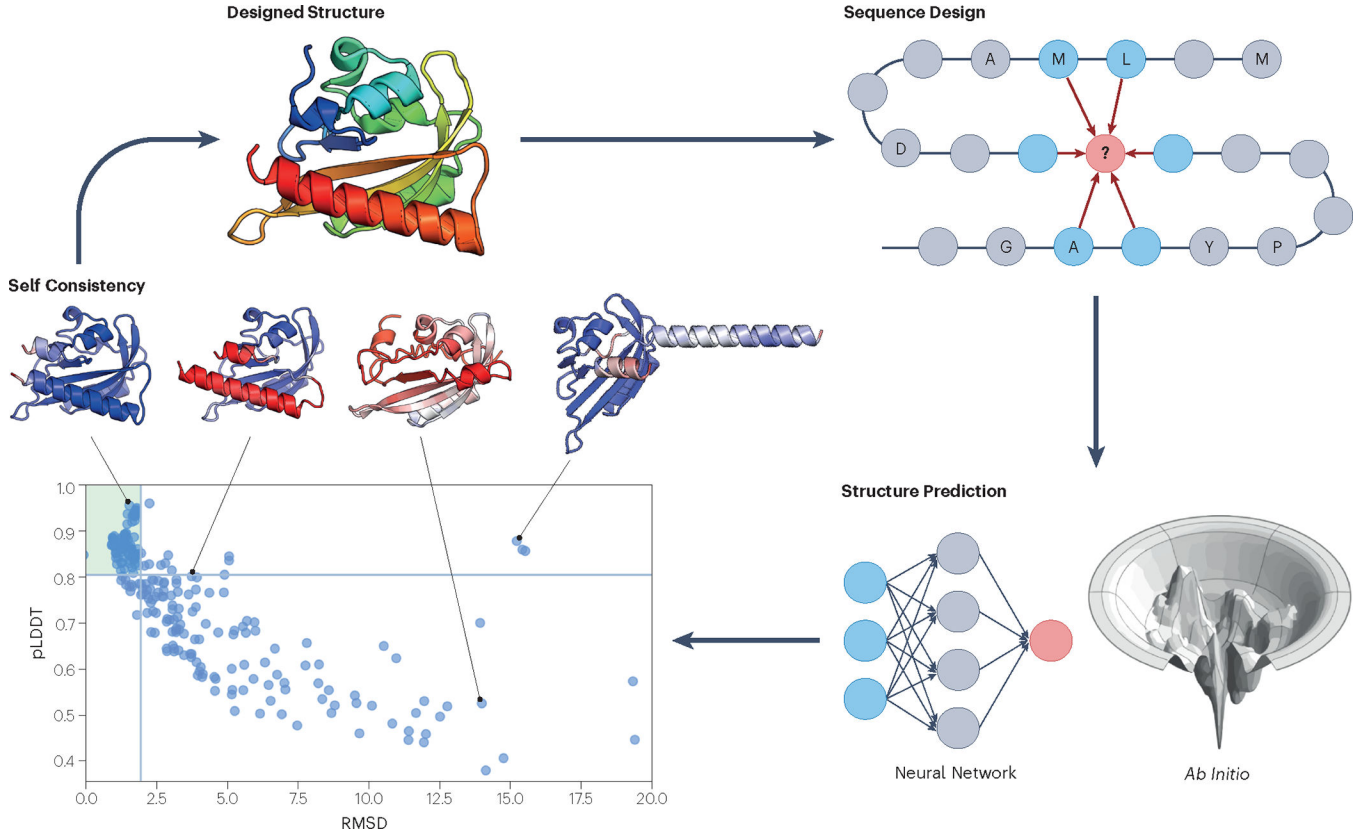

Fig. 5 |. Designing sequence to specify structure.

To complete the loop, designed sequences can be refolded with structure-prediction methods. Previously, ab initio structure prediction allowed for assessing the ‘folding landscape’ of a protein sequence and whether the designed structure is the lowest-energy conformation and is findable or foldable in this landscape. Now, with deep learning-based structure-prediction networks, accuracy and efficiency are much improved. Confidence metrics (pLDDT, pAE) and the root-mean-square deviation (RMSD) can be used as a scoring function and are predictive of experimental success. The folding funnel is by Ken Dill.

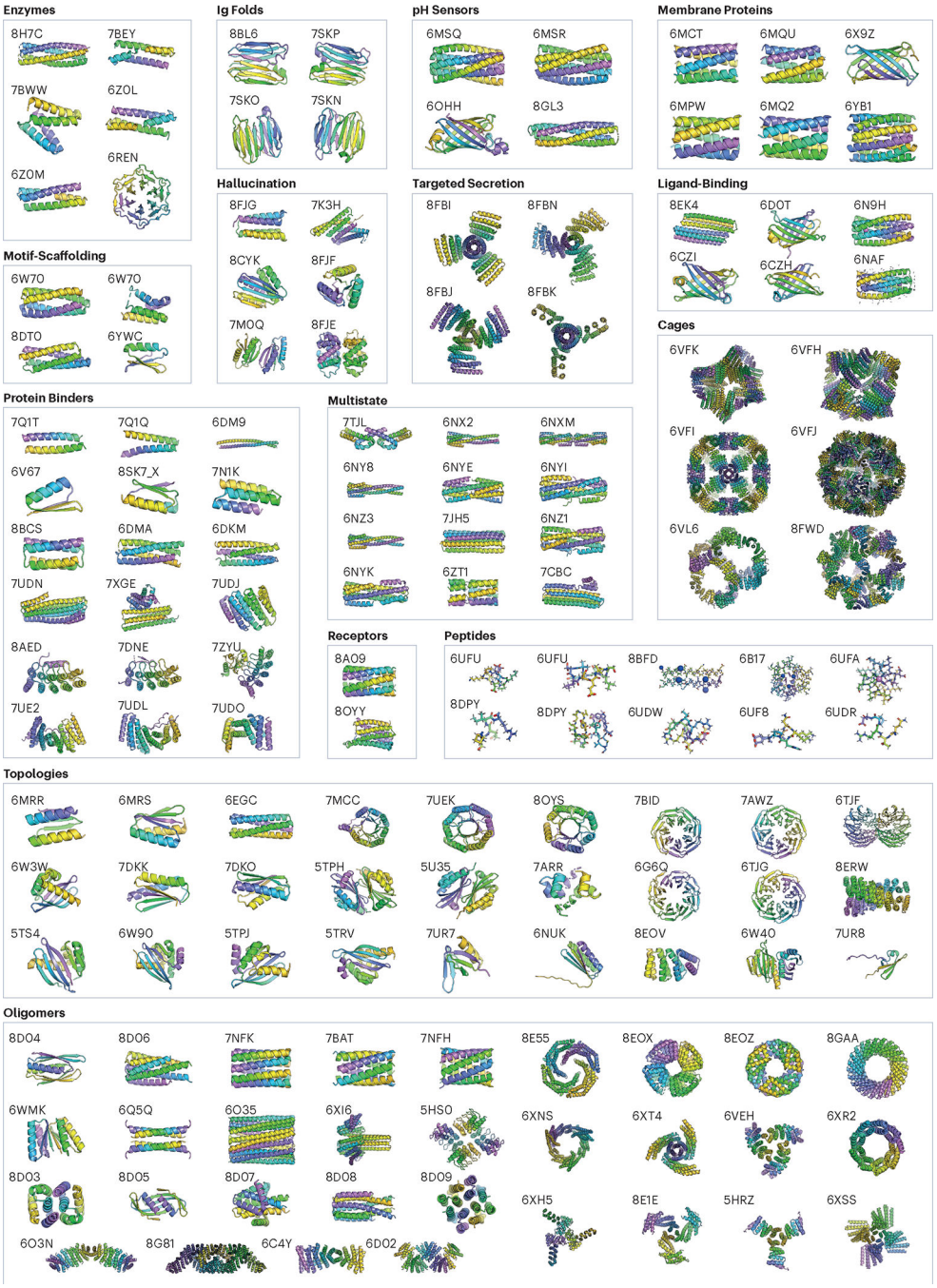

Fig. 6 |. Examples of functional de novo design.

Of particular note: 7K3H and 7M0Q are trRosetta-hallucinated structures[56]; 8CYK is an example of an adversarial sequence rescued by ProteinMPNN[103]; each of chains X, Y, Z in 8SK7 is a binder generated by RFdiffusion[20]; 6MSQ and 6MSR are designed by tuning contributions to the predicted free energy of folding from hydrophobic layers, pH-responsive polar layers and pH-independent polar layers[149]. The majority of structures are generated from a combination of methods, such as the various tools in Rosetta (8GAA)[150], library screening (8H7C)[151], rational design (7BEY)[152], parametric helical bundles (6MSQ)[149],

kinematic loop closure (6UD9)[153], rule-based design, rotamer interaction field (6D0T)[59], database fragment and/or interaction search and assembly (6W70, 6MCT)[61,97], customized docking protocols (8FWD) and negative design (6X9Z)[19]. Methods used to design select protein structures are listed in Supplementary Table 2. Ig, immunoglobulin.