peds
protein engineering design & selection

Protein Engineering, Design and Selection, 2021, vol. 00 no. 00, pp. 1–7
doi: 10.1093/protein/gzab007
Advance Access Publication Date: 00 Month 0000
Review

OXFORD

Review

Breakthroughs in computational design methods open up new frontiers for *de novo* protein engineering

Ben A. Meinen[1,2,3], and Christopher D. Bahl[1,2,3,*]

[1]Institute for Protein Innovation, Harvard Institutes of Medicine 4 Blackfan Circle, Room 941 Boston, MA 02115-5701 Boston, USA, [2]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA, USA, and [3]Department of Pediatrics, Harvard Medical School, Boston, MA, USA

*To whom correspondence should be addressed. E-mail: chris.bahl@proteininnovation.org

Edited By: Roberto Chica

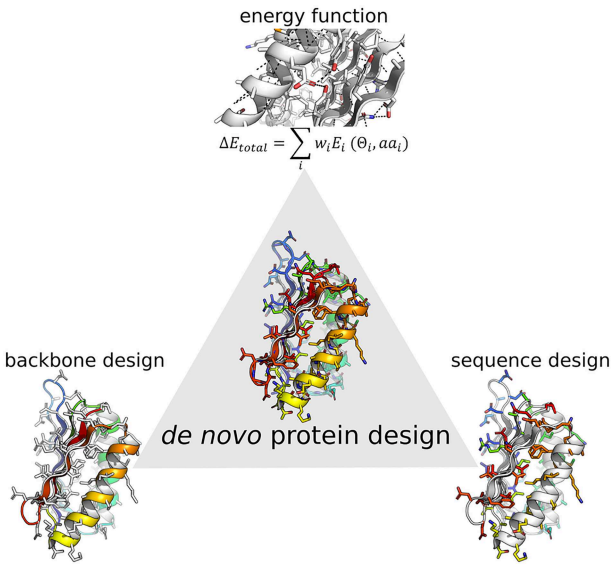
Received 30 November 2020; Revised 16 February 2021; Editorial Decision 8 March 2021; Accepted 8 March 2021

Abstract

Proteins catalyze the majority of chemical reactions in organisms, and harnessing this power has long been the focus of the protein engineering field. Computational protein design aims to create new proteins and functions *in silico*, and in doing so, accelerate the process, reduce costs and enable more sophisticated engineering goals to be accomplished. Challenges that very recently seemed impossible are now within reach thanks to several landmark advances in computational protein design methods. Here, we summarize these new methods, with a particular emphasis on *de novo* protein design advancements occurring within the past 5 years.

Key words: protein design, Rosetta, protein engineering, protein structure, energy, function

Graphical Abstract



Introduction

Creating a new protein that has a desired set of attributes and function with the click of a button is the dream of computational protein design. The ability to successfully engineer and design new proteins and protein functions is a testament to our increasing understanding of protein structure and folding. Although machine learning algorithms are starting to be successfully applied to protein design challenges, most advances in computational protein design have come from other areas such as improvements to the energy function or new protocols to sample the sequence design space.

Necessity often drives innovation, and *de novo* protein design, which is where a new protein is designed without using a natural template structure, poses unique challenges and has driven many substantial methodological advancements. The process of designing a protein *de novo* is commonly performed in two steps: generation of the protein mainchain tertiary structure, followed by design of the amino acid identity and side chain conformation at all residue positions (Fig. 1). The foundation for both design steps lies within the energy function that is used to evaluate the model. Each of these three essential aspects of computational protein design have their own challenges and have seen significant improvements in recent years. In this review, we summarize the major advances in computational protein design methods, with a particular emphasis on methods that enable *de novo* design.

The energy function drives protein design innovation

Computational protein design strives to find protein sequences that fold into low-energy, thermodynamically favorable structures. The energy function is the tool that calculates the energy of any given protein structure and distinguishes favorable from unfavorable structures. Physics-based energy functions such as Amber or CHARMM are commonly used for protein dynamics and modeling studies, but have rarely been used in *de novo* protein design applications (Suárez *et al.*, 2008; Rubenstein *et al.*, 2018). The OSPREY software package that has been used for protein redesign applications relies on an energy function that combines potentials form both Amber and CHARMM with additional terms (Gainza *et al.*, 2013; Lowegard *et al.*, 2020). ISAMBARD (Intelligent System for Analysis, Model Building and Rational Design) is a software package that can be used for parametric design of helical structures, and it uses its own force field based energy function called BUFF (Bristol University Docking Engine Force Field) (Wood *et al.*, 2017). Alternative approaches that only rely on statistical terms, such as dTERMen (Zhou *et al.*, 2020), have recently been developed and appear promising.

The most commonly used energy functions for protein design are from Rosetta, which contain physics-based energy terms as well as knowledge-based energy potentials derived from experimentally determined protein structures (Fig. 2A) (Alford *et al.*, 2017). The current default Rosetta energy function is REF2015, which combines many improvements made over the previous decades such as more accurate models of hydrogen bonding (O'Meara *et al.*, 2015), Van der Waals (Alford *et al.*, 2017), mainchain torsion angles (Leaver-Fay *et al.*, 2013) and sidechain rotamers (Shapovalov and Dunbrack, 2011). A significant improvement to the Rosetta energy function was made with the implementation of Rosetta-ECO and Rosetta-ICO (Pavlovicz *et al.*, 2020). This addition marks the first time that a Rosetta energy function has explicit consideration of solvent. These energy functions include additional energy terms for solvent

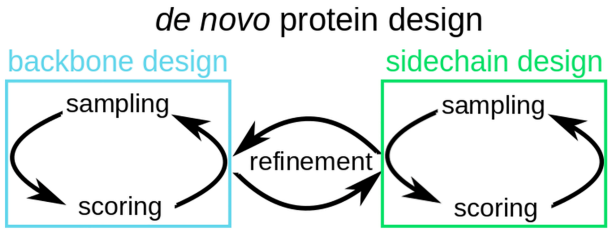


Fig. 1 *De novo* protein design is commonly performed in two interconnected steps. In both backbone and sidechain design steps the conformational and sequence space is sampled by different methods and algorithms and evaluated and scored by an energy function.

modeling and new sampling approaches that can simultaneously predict both sidechain geometry and coordinating waters. The addition of Rosetta-ICO/ECO improves the ability to identify favorable residue–residue or protein–protein interactions from unfavorable ones (Fig. 2C) (Pavlovicz *et al.*, 2020). Membrane proteins represent a unique challenge for protein design, and in the past, designers would simply deactivate the solvation energy term during membrane protein design (Lu *et al.*, 2018); this is a workaround rather than an optimal solution. Recently, the franklin2019 energy function was developed that contains an implicit membrane environment (Fig. 2D) (Alford *et al.*, 2020).

De novo protein design presents a unique challenge—the protein mainchain is usually generated prior to designing the amino acid sequence that will ultimately fit that mainchain configuration. The centroid and full-atom Rosetta energy functions require amino acid identities at each residue position in order to evaluate a protein structure. Thus, a critically important breakthrough was the development of an amino acid independent statistical potential for sidechain packing called motifscore (Fallas *et al.*, 2017) (Fig. 2B). Motfiscore leverages a precompiled database of aliphatic sidechain packing interactions to assign a score that indicates whether two given alpha carbons in a protein model will be able to interact favorably after performing full-atom sequence design. Thus, motifscore provides a means of assessing a protein tertiary structure in the absence of a primary amino acid sequence. By evaluating potential protein interface and core packing interactions based only on the mainchain, motifscore enables designers to more efficiently sample protein conformational landscapes prior to the computationally expensive full-atom design stage. Motifscore was first used to discover symmetric arrangements of proteins which could be designed in order to form homo-oligomers (Fallas *et al.*, 2017).

Protein sequence design

The grand challenge of computational protein design is to find an amino acid sequence that encodes for the desired protein function, structure, fold and stability. Fundamentally, protein sequence design is a very challenging search problem, as algorithms must be capable of sampling (at least) 20 different amino acids at each designable position, with each being able to adopt in a wide range of possible rotameric states. When performing amino acid sequence design on a native protein structure, it is common to avoid sampling substantial movement in the protein mainchain in order to reduce the risk of introducing mutations that will destabilize the native protein conformation. However, in *de novo* protein design, the protein mainchain generally requires refinement, and this refinement needs to be performed concurrently with amino acid sequence design.

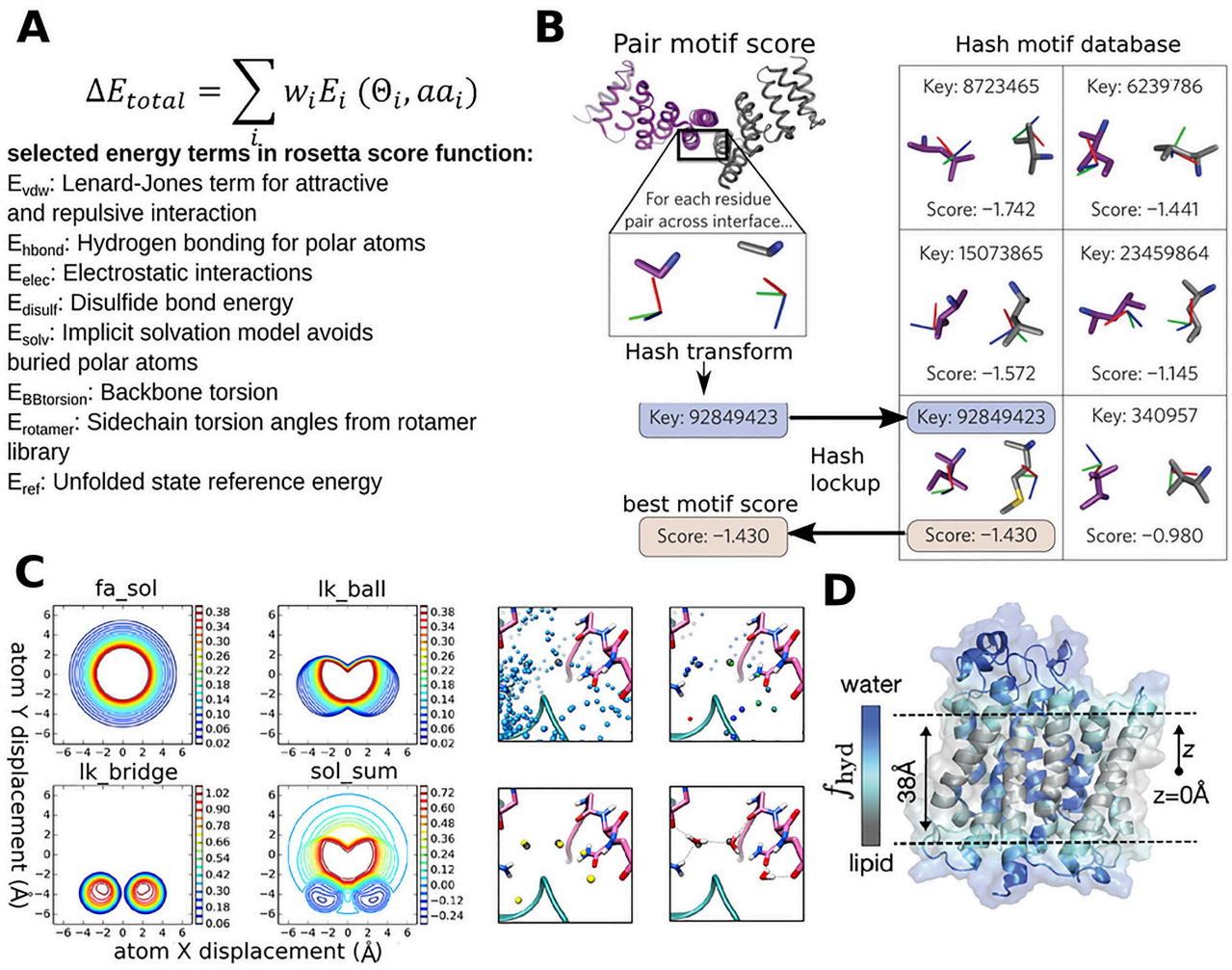


Fig. 2 Advancements in the Rosetta energy function. (A) The total energy in the Rosetta energy function is calculated from a linear combination of energy terms (E), that are scaled by an assigned weight (w), a term that describes the geometric degrees of freedom (Θ) and the chemical identities (aa) (Alford *et al.*, 2017). (B) Motifscore compares potential sidechain interactions to optimize designability of protein backbones (Fallas *et al.*, 2017). (C) Implicit and explicit water modeling introduced with Rosetta-ECO/ICO into the energy function (Pavlovicz *et al.*, 2020). (D) Introduction of implicit membrane models into Rosetta (Alford *et al.*, 2020).

The FastRelax protocol was originally developed to perform refinement of protein structures predicted *ab initio* (Tyka *et al.*, 2011); it was adapted to perform mainchain refinement and amino acid sequence design (a.k.a. 'flexible backbone' design) and renamed FastDesign (colloquially, this protocol is sometimes more accurately referred to as RelaxDesign) (Bhardwaj *et al.*, 2016). The original FastDesign protocol produces a bias for selecting small aliphatic amino acids in the protein core, which can result in 'collapse' of a protein model and lead to an alanine-rich core if mainchain motion is not constrained. Recently, the FastDesign algorithm was reparameterized, and the new protocol results in a more native-like distribution of protein residues in the core. Indeed, *de novo* miniproteins designed with this new FastDesign protocol exhibited improved Rosetta energies and were more stable *in vitro* (Maguire *et al.*, 2020).

Calculation of cooperative hydrogen bonding networks

The Rosetta energy function evaluates and estimates pairwise interactions. Although this enables the design of proteins with favorable hydrophobic packing, by itself, it cannot be used to design nonlocal structures such as extended networks of hydrogen bonds. The most successful approach for designing hydrogen bond networks is HBNet (Boyken *et al.*, 2016). This method discovers possible hydrogen-bond networks formed by polar side chains and identifies favorable networks that reduce the number of unsatisfied hydrogen-bond donors and acceptors. The efficacy of this approach was first demonstrated via the design symmetric homo-oligomeric proteins (Boyken *et al.*, 2016). The process of network enumeration becomes computationally expensive for larger systems, and the HBNet method was subsequently extended to incorporate sampling via Monte Carlo (Maguire *et al.*, 2018) (Fig. 3D). The ability to design larger connected and tunable hydrogen-bond networks has been used in multiple studies, including hetero-oligomers (Chen *et al.*, 2019), *de novo* designed proteins with dynamic structural behavior that can be used to build switchable logic gates (Boyken *et al.*, 2019; Langan *et al.*, 2019; Chen *et al.*, 2020), helical multipass transmembrane proteins (Lu *et al.*, 2018), and even transmembrane channels (Xu *et al.*, 2020).

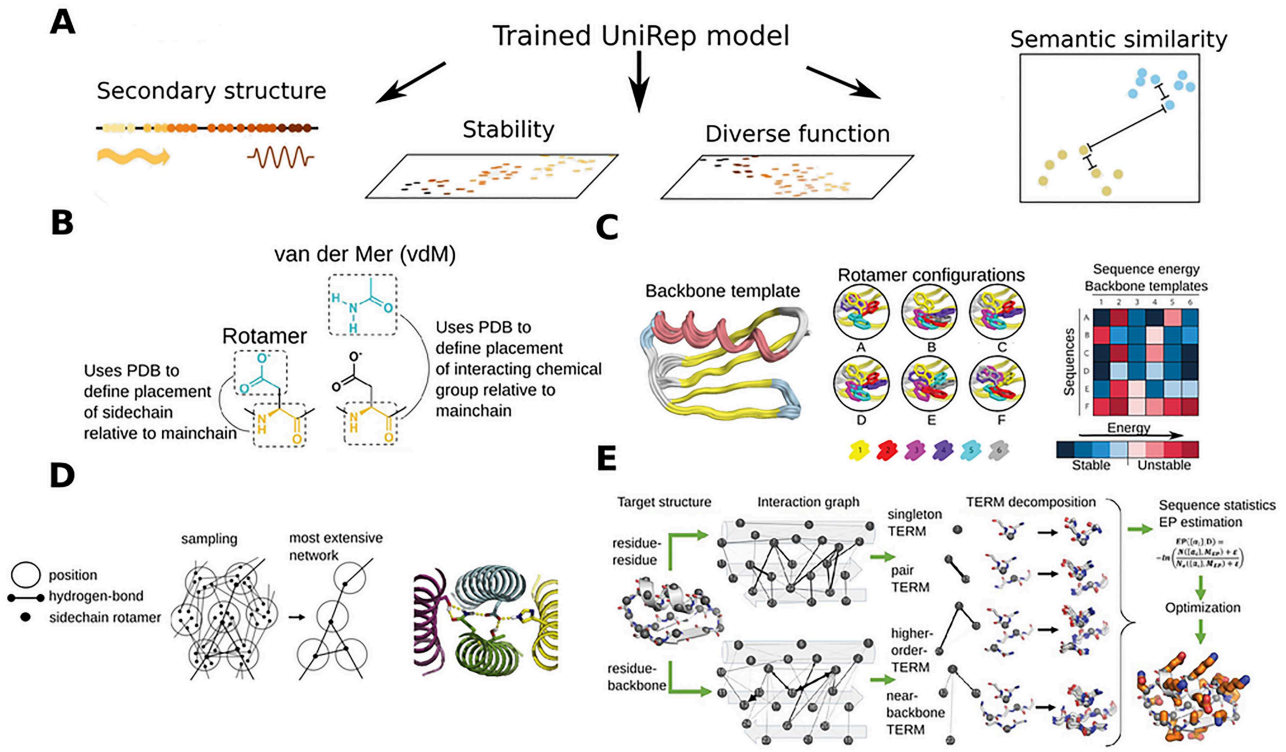


Fig. 3 Improvements in protein sequence design. (A) UniRep; supervised learning method that uses only protein sequences to improve protein properties or functions (Alley *et al.*, 2019). (B) Definition of a structural 'van der Mer' unit to design small molecule binding sites (Polizzi and DeGrado, 2020). (C) Multistate design (*meta*-MSD) finds sequences that occupy two states (Davey *et al.*, 2017). (D) HBNet samples low-energy hydrogen-bond networks that can be used for heterodimer design (Boyken *et al.*, 2016). (E) TERMs represent a data-driven approach to identify structural patterns and energies to predict and design protein structure (Zhou *et al.*, 2020).

Sampling approaches to incorporate binding function

The design of proteins that bind to other molecules, either small molecules or other macromolecules like proteins or nucleic acids, requires the precise positioning of amino acid side chains that will comprise the binding site. There are now several methods that provide generalizable solutions for accomplishing this. The Rotamer Interaction Field (RIF) dock method enables both small molecule (Dou *et al.*, 2018) and protein binding (Cao *et al.*, 2020) sites to be positioned. Another approach leverages a structural unit termed 'van der Mer' and searches for chemical groups and backbone coordination sites which can bind and coordinate a small molecule (Polizzi and DeGrado, 2020) (Fig. 3B). Finally, the Rosetta EnzDez protocol (Richter *et al.*, 2011) has been updated and improved recently to incorporate conformational ensembles; this now enables the design of small molecule biosensors (Glasgow *et al.*, 2019).

Multistate design

The rational design of proteins that occupy two or more stable conformations is a major challenge for the field, but some recent progress has been made toward this goal by employing multistate design approaches. In multistate design, the algorithm assesses the energy of sequence or structural perturbations on multiple protein structures at the same time (these are often different conformations of the same protein). One example is the Dancer proteins based on the small native Gβ1 protein (Fig. 3C). Dancer proteins can transition between two conformational states on a millisecond time scale (Davey *et al.*, 2017). To achieve this, a pool of unique structural templates were generated by changing the rotameric conformation of a central tryptophan residue, and Rosetta Backrub enabled sampling of conformations based on the changed rotamer. Then, a novel multistate design method called *meta*-MSD finds the lowest energy conformation for each template state and makes it possible to map the state onto an energy landscape. A second example is an all α-helical protein that can populate two distinct conformations; this was designed by first creating and validating two different proteins with distinct conformations and then searching for a sequence that satisfied both conformations (Wei *et al.*, 2020). In the future, finding sequences that can accommodate multiple states will be an important capability that enables the design of sophisticated biosensors and enzymes.

Informatic approaches for protein sequence design

Databases of protein sequences and structures have grown to enormous proportions in the last decade, and these data are increasingly being leveraged for protein design. Perhaps the most straightforward way to leverage informatics to augment protein design is to utilize multiple sequence alignments (MSAs) to guide amino acid selection. Two recent algorithms highlight the utility of this approach: PROSS and FuncLib. One longstanding challenge for computational design is protein stabilization, and PROSS utilizes MSAs to guide the Rosetta energy function to design mutations that enhance protein stability and expression levels (Goldenzweig *et al.*, 2016). FuncLib takes a similar approach, but focuses on mutations which enhance enzymatic activity (Khersonsky *et al.*, 2018). MSA based methods are limited to

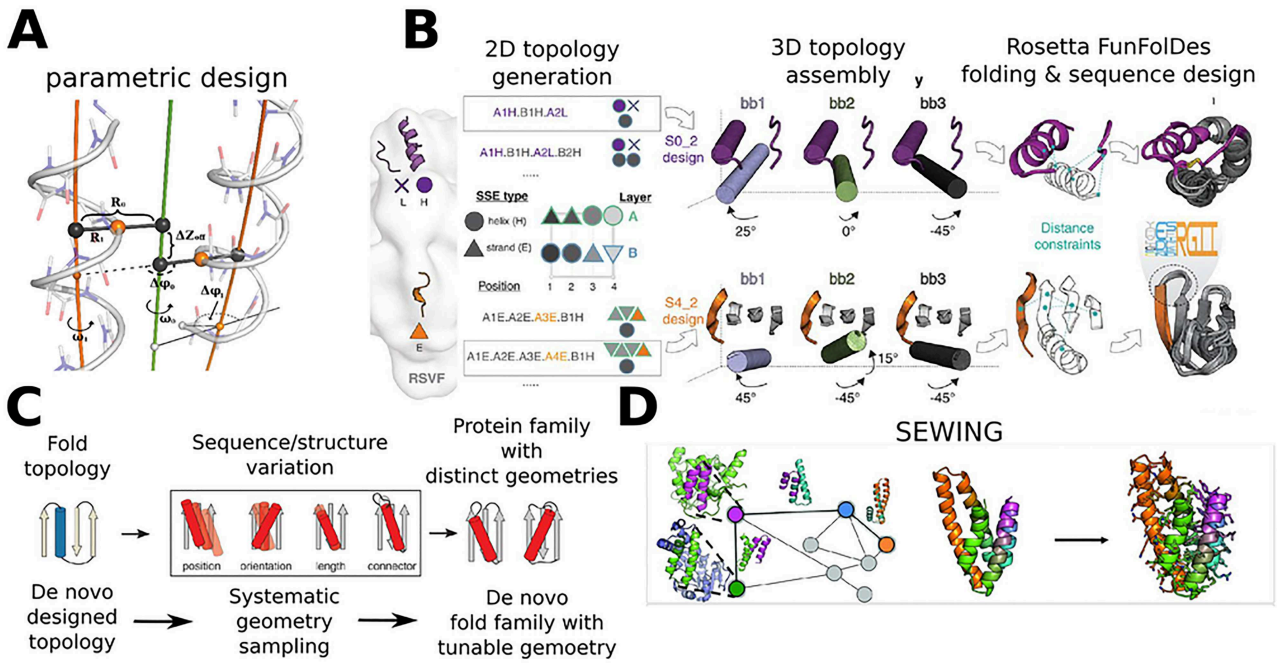


Fig. 4 Methods in protein backbone design. (A) Example of using geometrical expressions in parametric design to create *a*-helical bundles (Grigoryan and Degrado, 2011). (B) Topobuilder enables full control over protein topology (Sesterhenn *et al.*, 2020). (C) Sampling of loop–helix–loop (LUCS) fragments enables exploration of new geometries for defined topologies (Pan *et al.*, 2020). (D) SEWING builds proteins by combining fragments of larger protein structures allows high-fidelity assembly of *a*-helical protein backbones (Jacobs *et al.*, 2016).

proteins that have a sufficient number of homologous sequences that can be used to generate an MSA with a predictive signal (e.g. it can be challenging to use these approaches for highly conserved proteins that are specific to mammals).

Machine learning is a powerful way to utilize information encoded in protein sequences in a more general way than is possible via MSAs. Perhaps the most successful example of machine learning for protein design that does not leverage information about three dimensional protein structure is UniRep; this is a deep learning method that utilizes only protein sequences, and it can be used to rationally design mutations to improve protein properties or functions (Fig. 3A) (Alley *et al.*, 2019). Representing protein structure in a way that is amenable to machine learning is a significant challenge. One solution is to use local tertiary motifs, or TERMs. Zhou *et al.* used TERMs to develop statistical potentials derived from structures in the protein data bank and used them to generate a sequence-level pseudo energy table that can be used for structural analysis and design (Fig. 3E) (Zhou *et al.*, 2020). Together, these advances demonstrate that our current databases of protein sequences and structure possess enough information to enable data-driven approaches to protein design. In the future, methods that are able to merge classical physics-based energy functions with probabilistic energy potentials derived from machine learning are poised to significantly improve computational protein design.

Backbone design

When designing a protein, the amino acid sequence must direct the mainchain, or 'backbone' to adopt the designed conformation. Although the theoretical conformational space available to a protein is vast, only a small subset of conformations can be stably adopted. Thus, the biggest challenge to designing a protein backbone *de novo* is to generate a structure that is physically possible (Koga *et al.*, 2020). Successful approaches reduce the search space by defining constraints derived from parametric equations describing canonical protein structure (Fig. 4A) (Grigoryan and Degrado, 2011; Huang *et al.*, 2011; Thomson *et al.*, 2014) or by utilizing fragments of known protein structure (Koga *et al.*, 2012; Correia *et al.*, 2014; Jacobs *et al.*, 2016). Advances in parametric design methods have enabled the design of selective transmembrane proteins (Lu *et al.*, 2018; Xu *et al.*, 2020) and ion channels (Joh *et al.*, 2014), homo- and hetero-oligomers (Boyken *et al.*, 2016; Chen *et al.*, 2019), repeat proteins (Brunette *et al.*, 2015), miniproteins (Bhardwaj *et al.*, 2016) and antiapoptotic peptides that bind BCL-2 family proteins with high selectivity (Jenson *et al.*, 2018). Parametric design is most commonly used to design symmetric proteins, and there are limits to the types of function that can be imbued to these structures.

Fragment-based methods facilitate design of nonsymmetric structures, and there are three prominent methods: blueprint builder (Koga *et al.*, 2012), topology builder (Correia *et al.*, 2014) and structure extension (Jacobs *et al.*, 2016). At the vanguard of the *de novo* design renaissance is the blueprint builder method, which has been used to create proteins with a wide array of different shapes and structures (Lin *et al.*, 2015; Bhardwaj *et al.*, 2016; Marcos *et al.*, 2017, 2018; Rocklin *et al.*, 2017; Buchko *et al.*, 2018; Dou *et al.*, 2018), each explicitly specified by the user. One application of this approach has been to craft proteins that possess a ligand binding pocket (Basanta *et al.*, 2020), and these structures can be successfully functionalized to bind ligands (Bick *et al.*, 2017; Dou *et al.*, 2018). The design of hyperstable miniprotein affinity reagents is another example of successful functionalization of *de novo* protein scaffolds (Chevalier *et al.*, 2017; Cao *et al.*, 2020). Finally, it is possible to also use this approach to design 'protein mimics' that effectively scaffold a binding site from a natural protein onto a *de novo* structure (Silva *et al.*, 2019; Linsky *et al.*, 2020).

These more established approaches toward backbone generation are able to create idealized protein scaffolds, but they fail to explore the vast range of possible conformations and topologies. A recent and promising shift is to put function at the forefront and shape the protein backbone to provide structure and support. One approach to achieve this is Topobuilder (Bonet *et al.*, 2018; Yang *et al.*, 2021), which is an extension of the Fold From Loops protocol (Correia *et al.*, 2014). Topobuilder has been successfully used to design immunogens that can be used to induce neutralizing antibody responses by incorporating tertiary epitopes into a *de novo* protein (Sesterhenn *et al.*, 2020) (Fig. 4B). Another approach is structural extension with native-substructure graphs (SEWING) (Fig. 4D), which has successfully been used to create a range of proteins with novel structures as well as metal coordination sites (Jacobs *et al.*, 2016; Guffy *et al.*, 2018). Notably, a new method for creating backbone structures was recently developed that allows for the sampling of novel and tunable geometries in a predefined topology called loop–helix–loop unit combinatorial sampling (LUCS) (Pan *et al.*, 2020). This method also requires the manual specification of topology and insert points, but the protocol then samples loop–helix–loop fragments to close the insert points and samples geometries in an unbiased manner. The LUCS protocol has a high success rate for creating stable *de novo* proteins, and it was able to insert helical or loop fragments for ligand binding or protein–protein interactions into a stable scaffold (Pan *et al.*, 2020) (Fig. 4C). Excitingly, the pace of innovation is only accelerating, and entirely new methods for protein backbone design continue to be developed.

Summary and Outlook

The recent explosion of new protein design methods have greatly increased the breadth of challenges that can be solved computationally. For example, miniproteins are a therapeutically useful platform, but natural miniproteins are difficult to engineer using directed evolution (Correnti *et al.*, 2018). However, *de novo* designed miniproteins can readily be generated and functionalized to bind with high affinity to a desired epitope (Chevalier *et al.*, 2017; Crook *et al.*, 2017; Cao *et al.*, 2020).

Proteins are not static molecules, and the role that dynamics plays in determining protein structure and function is becoming increasingly clear. The next frontier for protein design lies in generalizable methods for predicting and designing dynamics; e.g. vibrations and coordinated motions. Indeed, recent studies have demonstrated that it is possible to rationally control protein dynamics in model systems (Davey *et al.*, 2017).

Finally, machine learning has revolutionized the field of protein structure prediction (Senior *et al.*, 2020), but thus far, it has had only minimal impact on the field of protein design. There is currently no generalizable solution that enables the design of protein backbone and sidechains simultaneously; this capability is necessary to precisely position amino acids that confer protein function. Machine learning will almost certainly play a major role in solving this challenge, and in doing so, unlock the full potential of protein engineering.

References

Alford,R.F., Fleming,P.J., Fleming,K.G. and Gray,J.J. (2020) *Biophys. J.*, **118**, 2042–2055.

Alford,R.F., Leaver-Fay,A., Jeliazkov,J.R. *et al.* (2017) *J. Chem. Theory Comput.*, **13**, 3031–3048.

Alley,E.C., Khimulya,G., Biswas,S., AlQuraishi,M. and Church,G.M. (2019) *Nat. Methods*, **16**, 1315–1322.

Basanta,B., Bick,M.J., Bera,A.K., Norn,C., Chow,C.M., Carter,L.P., Goreshnik,I., Dimaio,F. and Baker,D. (2020) *Proc. Natl. Acad. Sci.*, **117**, 22135–22145.

Bhardwaj,G., Mulligan,V.K., Bahl,C.D. *et al.* (2016) *Nature*, **538**, 329–335.

Bick,M.J., Greisen,P.J., Morey,K.J. *et al.* (2017) *ELife*, **6**, e28909

Bonet,J., Wehrle,S., Schriever,K. *et al.* (2018) *PLOS Computational Biology*, **14**, e1006623.

Boyken,S.E., Benhaim,M.A., Busch,F. *et al.* (2019) *Science*, **364**, 658–664.

Boyken,S.E., Chen,Z., Groves,B. and Langan,R.A. (2016) *Science*, **352**, 680–687.

Brunette,T.J., Parmeggiani,F., Huang,P.S., Bhabha,G., Ekiert,D.C., Tsutakawa,S.E., Hura,G.L., Tainer,J.A. and Baker,D. (2015) *Nature*, **528**, 580–584.

Buchko,G.W., Pulavarti,S.V.S.R.K., Ovchinnikov,V. *et al.* (2018) *Protein Sci.*, **27**, 1611–1623.

Cao,L., Goreshnik,I., Coventry,B. *et al.* (2020) *Science*, **370**:426–431.

Chen,Z., Boyken,S.E., Jia,M. *et al.* (2019) *Nature*, **565**, 106–111.

Chen,Z., Kibler,R.D., Hunt,A. *et al.* (2020) *Science*, **368**, 78–84.

Chevalier,A., Silva,D.A., Rocklin,G.J. *et al.* (2017) *Nature*, **550**, 74–79.

Correia,B.E., Bates,J.T., Loomis,R.J. *et al.* (2014) *Nature*, **507**, 201–206.

Correnti,C.E., Gewe,M.M., Mehlin,C. *et al.* (2018) *Nat. Struct. Mol. Biol.*, **25**, 270–278.

Crook,Z.R., Sevilla,G.P., Friend,D. *et al.* (2017) *Nat. Commun.*, **8**, 2244.

Davey,J.A., Damry,A.M., Goto,N.K. and Chica,R.A. (2017) *Nat. Chem. Biol.*, **13**, 1280–1285.

Dou,J., Vorobieva,A.A., Sheffler,W. *et al.* (2018) *Nature*, **561**, 485–491.

Fallas,J.A., Ueda,G., Sheffler,W. *et al.* (2017) *Nat. Chem.*, **9**, 353–360.

Gainza,P., Roberts,K.E., Georgiev,I. *et al.* (2013) *Methods Enzymol.*, **523**, 87–107.

Glasgow,A.A., Huang,Y.M., Mandell,D.J. *et al.* (2019) *Science*, **366**, 1024–1028.

Goldenzweig,A., Goldsmith,M., Hill,S.E. *et al.* (2016) *Mol. Cell*, **63**, 337–346.

Grigoryan,G. and Degrado,W.F. (2011) *J. Mol. Biol.*, **405**, 1079–1100.

Guffy,S.L., Teets,F.D., Langlois,M.I. and Kuhlman,B. (2018) *J. Chem. Inf. Model.*, **58**, 895–901.

Huang,P.S., Ban,Y.E.A., Richter,F., Andre,I., Vernon,R., Schief,W.R. and Baker,D. RosettaRemodel: a generalized framework for flexible backbone protein design. (2011) *PLoS One*, **6**, e24109.

Jacobs,T.M., Williams,B., Williams,T., Xu,X., Eletsky,A., Federizon,J.F., Szyperski,T. and Kuhlman,B. (2016) *Science*, **352**, 687–690.

Jenson,J.M., Xue,V., Stretz,L., Mandal,T., Reich,L.L. and Keating,A.E. (2018) *Proc. Natl. Acad. Sci. U. S. A.*, **115**, E10342–E10351.

Joh,N.H., Wang,T., Bhate,M.P., Acharya,R., Wu,Y., Grabe,M., Hong,M., Grigoryan,G. and DeGrado,W.F. (2014) *Science*, **346**, 1520–1524.

Khersonsky,O., Lipsh,R., Avizemer,Z. *et al.* (2018) *Mol. Cell*, **72**, 178, e5–186.

Koga,N., Tatsumi-Koga,R., Liu,G., Xiao,R., Acton,T.B., Montelione,G.T. and Baker,D. (2012) *Nature*, **491**, 222–227.

Koga,R., Yamamoto,M., Kosugi,T., Kobayashi,N., Sugiki,T., Fujiwara,T. and Koga,N. (2020) *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 31149–31156.

Langan,R.A., Boyken,S.E., Ng,A.H. *et al.* (2019) *Nature*, **572**, 205–210.

Leaver-Fay,A., O'Meara,M.J., Tyka,M. *et al.* (2013) *Methods Enzymol.*, **523**, 109–143.

Linsky,T.W., Vergara,R., Codina,N. *et al.* (2020) *Science.*, **370**, 1208–1214.

Lin,Y.R., Koga,N., Tatsumi-Koga,R., Liu,G., Clouser,A.F., Montelione,G.T. and Baker,D. (2015) *Proc. Natl. Acad. Sci. U. S. A.*, **112**, E5478–E5485.

Lowegard,A.U., Frenkel,M.S., Holt,G.T., Jou,J.D., Ojewole,A.A. and Donald,B.R. (2020) *PLoS Comput. Biol.*, **16**, e1007447.

Lu,P., Min,D., DiMaio,F. *et al.* (2018) *Science*, **359**, 1042–1046.

Maguire,J.B., Boyken,S.E., Baker,D. and Kuhlman,B. (2018) *J. Chem. Theory Comput.*, **14**, 2751–2760.

Maguire,J., Haddox,H., Strickland,D. *et al.* (2020) *Front. Bioeng.Biotechnol.*, **89**, 436–449.

Marcos,E., Basanta,B., Chidyausiku,T.M. *et al.* (2017) *Science*, **355**, 201–206.

Marcos,E., Chidyausiku,T.M., McShan,A.C. *et al.* (2018) *Nat. Struct. Mol. Biol.*, **25**, 1028–1034.

O'Meara,M.J., Leaver-Fay,A., Tyka,M.D. *et al.* (2015) *J. Chem. Theory Comput.*, **11**, 609–622.

Pan,X., Thompson,M.C., Zhang,Y., Liu,L., Fraser,J.S., Kelly,M.J.S. and Kortemme,T. (2020) *Science*, **369**, 1132–1136.

Pavlovicz,R.E., Park,H. and DiMaio,F. (2020) *PLoS Comput. Biol.*, **16**, e1008103.

Polizzi,N.F. and DeGrado,W.F. (2020) *Science*, **369**, 1227–1233.

Richter,F., Leaver-Fay,A., Khare,S.D., Bjelic,S. and Baker,D. (2011) *PLoS One*, **6**, e19230.

Rocklin,G.J., Chidyausiku,T.M., Goreshnik,I. *et al.* (2017) *Science*, **357**, 168–175.

Rubenstein,A.B., Blacklock,K., Nguyen,H., Case,D.A. and Khare,S.D. (2018) *J. Chem. Theory Comput.*, **14**, 6015–6025.

Senior,A.W., Evans,R., Jumper,J. *et al.* (2020) *Nature*, **577**, 706–710.

Sesterhenn,F., Yang,C., Bonet,J. *et al.* (2020) *Science*, **368**, , eaay5051.

Shapovalov,M.V. and Dunbrack,R.L.Jr. (2011) *Structure*, **19**, 844–858.

Silva,D.A., Yu,S., Ulge,U.Y. *et al.* (2019) *Nature*, **565**, 186–191.

Suárez,M., Tortosa,P. and Jaramillo,A. (2008) *Syst. Synth. Biol.*, **2**, 105–113.

Thomson,A.R., Wood,C.W., Burton,A.J., Bartlett,G.J., Sessions,R.B., Leo Brady,R. and Woolfson,D.N. (2014) *Science*, **346**, 485–488.

Tyka,M.D., Keedy,D.A., André,I., Dimaio,F., Song,Y., Richardson,D.C., Richardson,J.S. and Baker,D. (2011) *J. Mol. Biol.*, **405**, 607–618.

Wei,K.Y., Moschidi,D., Bick,M.J. *et al.* (2020) *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 7208–7215.

Wood,C.W., Heal,J.W., Thomson,A.R., Bartlett,G.J., Ibarra,A.Á., Brady,R.L., Sessions,R.B. and Woolfson,D.N. (2017) *Bioinformatics*, **33**, 3043–3050.

Xu,C., Lu,P., Gamal El-Din,T.M. *et al.* (2020) *Nature*, **585**, 129–134.

Yang,C., Sesterhenn,F., Bonet,J. *et al.* (2021) *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-020-00699-x

Zhou,J., Panaitiu,A.E. and Grigoryan,G. (2020) *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 1059–1068.