nature biotechnology

Review article

https://doi.org/10.1038/s41587-024-02127-0

Machine learning for functional protein design

Received: 1 July 2023

Accepted: 5 January 2024

Published online: 15 February 2024

Check for updates

Pascal Notin[1,2,5] ✉, Nathan Rollins[3,5] ✉, Yarin Gal[2], Chris Sander[1,4] & Debora Marks[1,4] ✉

Recent breakthroughs in AI coupled with the rapid accumulation of protein sequence and structure data have radically transformed computational protein design. New methods promise to escape the constraints of natural and laboratory evolution, accelerating the generation of proteins for applications in biotechnology and medicine. To make sense of the exploding diversity of machine learning approaches, we introduce a unifying framework that classifies models on the basis of their use of three core data modalities: sequences, structures and functional labels. We discuss the new capabilities and outstanding challenges for the practical design of enzymes, antibodies, vaccines, nanomachines and more. We then highlight trends shaping the future of this field, from large-scale assays to more robust benchmarks, multimodal foundation models, enhanced sampling strategies and laboratory automation.

Proteins fulfill a wide range of functions in nature, with that functional diversity encoded in their amino acid sequences. The goal of protein design is to create new proteins by discovering sequences with functions that enhance or extend beyond those of existing proteins—an aim with the potential to address globally pressing problems in healthcare, agriculture and sustainability. However, the potential design space is massive and sparsely functional: there are more unique sequences of 100 amino acids than the number of atoms in the universe, and only a small fraction of these has desired functions in the context of interest (for example, organism, temperature, pH). The quantitative map of the representations of proteins to their functions is referred to as a 'fitness landscape'. Given the impossibility to exhaustively list all possible amino acid combinations, let alone quantify their properties experimentally or computationally in different contexts, one of the first challenges that protein design is faced with is narrowing the search within the fitness landscape to a tractable space. A multitude of strategies have been developed to address this challenge: from rational design methods that select the most promising mutants on the basis of a deep understanding of a given protein structure and function, to experimental methods testing a broader range of variants (for example, directed evolution, combinatorial libraries), to biophysics-based models of protein structure, folding and interactions—a staple of computational design.

Machine learning methods have recently emerged as another strategy to efficiently explore the functional protein space, given their ability to learn complex distributions that model fitness landscapes from data. This ability typically increases with the quantity and quality of data available for training, as well as the aptness of the underlying algorithms to learn from these data via the right inductive biases—that is, the set of assumptions or constraints encoded in the model architecture. The massive progress in DNA sequencing over the past two decades, combined with improvements in the experimental determination of protein structure and properties, has provided the required foundational data for machine learning for protein design to be successful (Box 1). In parallel, algorithmic and computing advances have led to an increasing capacity to model distributions over these various data inputs, leading to a broad collection of performant protein design models achieving diverse objectives (Box 2).

While in practice it may be effective to combine several design strategies (for example, generate initial designs with a machine learning model and then optimize with a biophysical method), we focus here on machine learning-based design methods. After providing an overview of the breadth of tasks in protein design, we review the fundamental modeling approaches involved and discuss their practical applications, successes and limitations. Given the rapid pace of progress

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [2]Department of Computer Science, University of Oxford, Oxford, UK. [3]Seismic Therapeutic, Cambridge, MA, USA. [4]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [5]These authors contributed equally: Pascal Notin, Nathan Rollins. ✉e-mail: pascal_notin@hms.harvard.edu; nrollins.home@gmail.com; debbie@hms.harvard.edu

BOX 1

The three core modalities of machine learning for functional protein design

Protein design models are trained on a combination of sequences, structures, and functional labels. Each modality has unique virtues and caveats, depending on the abundance of data, the need for human knowledge and intervention, and the proximity of data examples to the desired function.

Sequences

- **Billions of publicly available sequences** encode evolutionary constraints, encompassing the diversity of protein families and mitigating the need to generate data in the lab.
- Models of evolved sequences can optimize for **functions that are selected for during evolution**.

Structures

- **Thousands of publicly available structures** provide 3D detail of the biochemical interactions that underpin protein folding and function.
- Models of sequences and structures can optimize for **custom folds, active sites**, and **binding complexes**, but often require expert knowledge of key residues and interactions.

Functional labels

- **Lab-generated functional labels** explore new binding targets, reactions and biochemical conditions. Applicable datasets are often sparse and bespoke, but the ability to generate larger and more generalizable libraries is increasing.
- Models of sequences and labels can optimize for **new functions** or for **existing functions in new conditions**.

in this field, our emphasis is on teasing out the overarching principles that characterize all methods, rather than focusing on the details of the current top-performing models that are sure to be outperformed in the near future. We conclude by providing a vision for the field of protein engineering as different threads of research are converging.

Objectives of machine learning for functional protein design

The design objectives supported by machine learning can be broadly classified into three groups, depending on whether we start from a known protein or from scratch and, for known proteins, whether we enhance its existing function or create a new function (Fig. 1). We review these different strategies and connect them with the machine learning approaches discussed in Box 2.

Redesign to enhance an existing function

The goal of protein enhancement is to start from a protein (natural or otherwise) that already possesses the desired function and introduce mutations to improve its properties or achieve the original biological function in different conditions. The aim could be to enhance the main function of the protein (for example, catalytic activity, binding affinity to a specific target), to enhance another of its attributes (for example, thermostability[1–4]) or to mitigate undesirable interactions with other molecules, such as reducing the immunogenicity of therapeutic proteins by altering epitopes[5–11]. One way to enhance intrinsic properties such as stability is to sample high-probability sequences from a sequence-based model[2,3] or a sequence–structure model[1,4]. Sequence-based models stand out for their ability to accurately predict effects of mutations on diverse evolutionary phenotypes (binding, stability, enzymatic function and more)[12–18] and to produce designs achieving the mixture of phenotypes key to function[2,3,19–21], owing to the depth of sequence data available for many protein families. Sequence–structure models such as inverse folding models can generate highly stable protein sequences[1,4,22,23]. However, they have so far relied on protein family sequence profiles and simulated residue-interaction fields to account for constraints on functions such as enzymatic activity[22,23]. Circumstantially, when sufficient mutation–phenotype data exist for the property of interest, sequence–label models can also prove useful to guide design, for example, for intended function[24,25], stability[26–28] or immune epitope[5–11] prediction.

Redesign for a new function

The objective here is to design a protein with a new function by working from an existing protein with a related function (for example, shifting a binder or enzyme to act on a new target[29,30]). This requires either a detailed understanding of the mechanism of function or ample data relating sequence to the new function. Consequently, most approaches have relied on sequence–label models. These data can be obtained by selecting sequences according to the reaction of interest via measured phenotypes of natural proteins[31], deep mutational scans[25], next-generation sequencing of library selection[32] or directed evolution experiments[29,30,33–35]. Sequence-based models can also be used to generate libraries of 'new family members' to pan for secondary properties[3,19–21,36–38] (for example, enzyme functionality within *Escherichia coli*[36], viral gene delivery to a tissue[37] or antibody binding affinity to a target[14,39,40]). This panning may cost more than screening a small label-driven library when sufficient labels exist for training, but can reduce costs (for example, necessary iterations and scale of selection) compared to random libraries by enriching for sequences with good intrinsic properties and high diversity[14,37]. Sequence–structure models can also be leveraged for this objective, for instance, to redesign a region of a protein to achieve a new binding interaction or to insert an active site, as has been achieved repeatedly with non-machine learning methods[41–43]. However, application of sequence–structure models has so far leaned toward de novo objectives, and, even when templating from existing proteins, the templates have often been abstracted to fragments and topology constraints[44].

De novo design

Machine learning-based design of sequences with de novo folds focuses on sequence–structure models. These methods can generate sequences with diverse 3D folds and multimer arrangements with a high success rate of stable expression[22,23,45–47]. The motivation to design sequences based on 3D structure results from the critical role of structure in our understanding of protein function. The 3D structure of a protein enables us to make assertions about physicochemical interactions and is a convenient representation for inferring or specifying constraints on function. De novo design requires function constraints on sequence and structure, often derived from other existing proteins, such as metal-binding sites and fragments of protein–protein complexes[23,48]. Excitingly, it is becoming possible to design a de novo protein on the basis of the target structure alone for both protein–protein binding[43,48,49] and small-molecule binding[44,49,50]. Recently, luciferase enzymatic activity has been achieved by a de novo protein, with model-generated 3D structure and sequence, albeit templated on an existing family of small-molecule-binding proteins[44]. Although some applications require thousands of designs to be assayed, the capabilities of de novo design are growing rapidly. Lastly, it is possible to use sequence-based models to generate new proteins not explicitly templated on an existing protein and even with predicted folds unique from those in the PDB[51,52]. However, to achieve a new desired function

BOX 2

Types of machine learning models for protein design

The majority of machine learning models for protein design can be broadly categorized into three groups, based on the way proteins are represented, the data used for training and the probability that the underlying algorithm seeks to learn (Fig. 3). We use a unifying probabilistic framework to facilitate comparisons between the different methods in this section and in Fig. 3. While certain models do not explicitly learn a distribution, one can always cast the corresponding tasks as implicitly modeling a distribution, adopting a Bayesian viewpoint.

Sequence-based models. This model class can be split into two distinct groups. The first group, sequence-only models, learns a generative model $P(x)$ of the primary structure x of a given protein. By training on a large collection of protein sequences, they aim to implicitly capture the biochemical constraints that characterize the proteins present in the training set. Models in that category were classically 'family-specific' alignment-based models and trained on a set of homologous sequences contained in a multiple-sequence alignment. While initial alignment-based models focused on position-specific predictions independent from other sequence residues (for example, position-specific scoring matrices), subsequent models considered pairs of residues[12,174] and, more recently, were generative models of the full sequence[13,14,16,38]. Building on the intuition that certain amino acid constraints or patterns may generalize across protein families, 'family-agnostic' protein language models—trained on unaligned sequences across protein families—then emerged as a practical alternative covering all proteins with a single model. This has led to a wide diversity of models, inspired by learning paradigms initially introduced in the natural language-processing literature, such as autoregressive modeling[20,52,61,141], masked-language modeling[15,146,175,176] or seq2seq architectures[177,178]. However, without relying more explicitly on homology (for example, via fine-tuning on alignments[15]), family-agnostic models have not been able to match the fitness-prediction abilities of the best family-specific model[18,136,137,144,169]. This observation subsequently gave rise to a multitude of hybrid models that sought to combine the relative strengths of each approach[144,145,179,180]. Recently, diffusion models of single-sequence or MSA inputs have also been proposed as a promising avenue for learning the process to generate full protein sequences from scratch[181]. The second group, conditional sequence models, further condition the generative process $P(x|t)$ on broad taxonomic groups or gene ontology annotations t to provide more control over the nature and properties of generated sequences. Several architectures have been proposed based on autoregressive modeling[20,153] or masked-language modeling[182].

Sequence–label models. When a sufficiently large number of labels for the property of interest are available, it becomes possible to train discriminative supervised models learning a distribution $P(y|x)$ of a functional label y for a given input sequence x. Functional labels are typically measurements collected in the laboratory via massively parallel sequencing of directed evolution campaigns[29,30,35,40], mutation libraries[25,27,37,89,183,184], nature sourced (for example, antibody repertoires[32], chimeric libraries[34]), or partial measurements that supplement natural sequence data[24,31]. The sequence representation can be either a simple one-hot encoding, physicochemical properties and other handcrafted features or, increasingly, embeddings extracted from sequence-based models, giving rise to label-efficient semi-supervised architectures. The trained regressor is usually lightweight to avoid overfitting, for example, a ridge regression, Gaussian process, shallow CNN or dense network[159,185–187]. Discriminative models provide an efficient way to predict phenotypical values for a large list of potential designs and prioritize the most promising candidates. However, the quality of these candidates directly depends on the quality of the external procedure (for example, combinatorial libraries, sampling with a separate unsupervised model) that was used to craft the initial list of prioritized mutants to assess. Conversely, label-conditioned generative models, such as conditional variational autoencoders (VAEs)[29], guided diffusion[188], Regression Transformer[154] or ProteinNPT[155], learn an approximation of the joint probability $P(x,y)$ or the conditional probability $P(x|y)$ for a sequence x and functional label y. They enable the generation of new sequences conditioned on a desired phenotypical value, leading to potentially more effective end-to-end procedures. Lastly, while the majority of supervised models have relied on representations of the primary structure (one dimensional) of proteins of interest, some architectures are instead based on their tertiary 3D structure[34,189] and would be more adequately characterized as structure–label models.

Structure-based models. There are four main categories of structure-based models that may be used for protein design. Structure prediction models[143,146,190,191] seek to predict the tertiary structure z of a protein on the basis of its primary structure x. Structure generation models, based on generative adversarial networks[192], VAEs[96] or, more recently, diffusion models[23,77,193], are trained to directly learn the probability $P(z)$. Inverse folding models[22,62,107,194,195] learn the probability $P(x|z)$ of a protein sequence x, conditioned on a 3D structure z, where the structure is typically encoded with a graph neural network[196–199]. Lastly, holistic design approaches, such as protein hallucination[45,48,200], inpainting[48] or ProteinGenerator[100], learn to model the joint probability $P(x,z)$ of the sequence x and structure z. In the text, we often refer to sequence–structure models, which encompass both inverse folding as well as the joint sequence and structure models. To eventually produce new sequence designs, structure generation models must be paired with these sequence–structure models. In that case, the combined architecture can itself be seen as a joint model of sequence and structure since, using the chain rules for probabilities, $P(x,z)=P(x|z)P(z)$.

Choosing a model architecture. The main drivers in the selection of a particular model are the desired design objective (as discussed in Objectives of machine learning for functional protein design) and available data to support that goal (for example, a sufficiently large number of labels to train sequence–label models). Another consideration is how the model will be effectively used in practice. Generative models, whether they are sequence or structure based, provide a way to sample new proteins that resemble the data they have been trained on. A sound sampling process producing natural-like proteins coupled with a robust experimental pipeline to measure the actual properties of generated objects is key. Additional controls (for example, taxonomic labels[20], enzyme classification[153]) with which we can condition the sampling process may help in increasing the usefulness of each sample. Alternatively, sequence–label architectures provide diverse ways

(continued from previous page)

to prioritize a subset of variants for subsequent experimental validation. For instance, if the model outputs both predictions along with the corresponding uncertainty (for example, Gaussian process, Bayesian neural network), one can frame the iterative design approach under a batch Bayesian optimization framework. If the function mapping the protein representation to the property of interest is differentiable (for example, when jointly training a regressor with a VAE[201–203] or in guided diffusion[188]), one can instead use gradient ascent. Lastly, as long as enough labels are available to train them, supervised generative models[154,155] provide finer-grained control over the sampling process than their unsupervised counterparts.

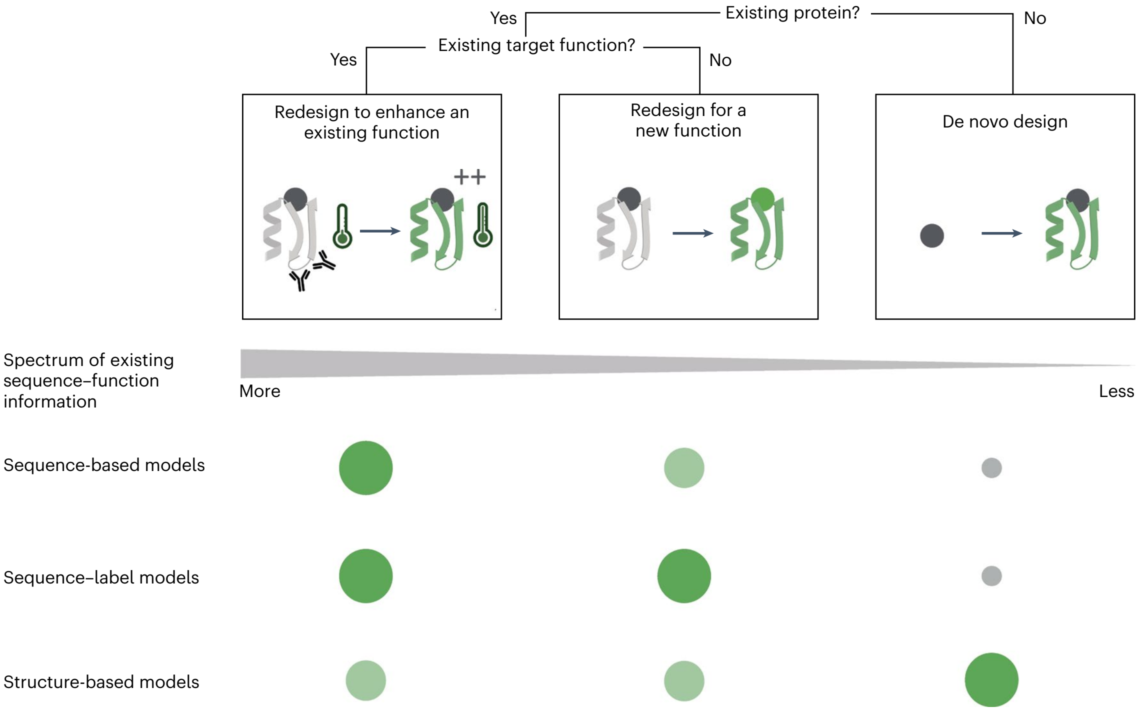

Existing protein?
Yes
No
Existing target function?
Yes
No
Redesign to enhance an existing function
++
Redesign for a new function
De novo design
Spectrum of existing sequence–function information
More
Less
Sequence-based models
Sequence–label models
Structure-based models

Fig. 1 | Protein design objectives. The various protein design objectives can be placed into three groups based on how far the design target is from existing functional proteins. For each group, different model classes may be more appropriate depending on available sequence–function information. The size of the circle indicates how suited a given approach is to a particular design task. When selecting a design method, one can decide to ignore the existence of relevant functional proteins in nature. For instance, even if a known protein already carries out the desired function, de novo design methods can be used to delve into regions of sequence space not previously explored by evolution but yielding the same function.

in the absence of a template protein, it is foreseeably more practical to incorporate structural constraints, for instance, via hybrid models[46].

Redesign versus de novo design

The common distinction in protein design between redesign (creating new sequences based on existing proteins) and de novo design (creating new sequences based on new folds) is more nuanced than the dichotomy implies. It is more accurately represented as a spectrum of strategies (Fig. 1), all of which leverage function from natural sequence and structural elements to varying degrees. Even de novo designs, despite their seemingly new sequence or overall structure, are products of training data from natural sequences and structures and often incorporate functional motifs from existing proteins. Therefore, when faced with a new protein design challenge, the initial question should be: what existing protein with similar function can I use as a template? The extent of the match between the function ask and the template protein then determines the sources of data and the model strategy that are best suited to the task.

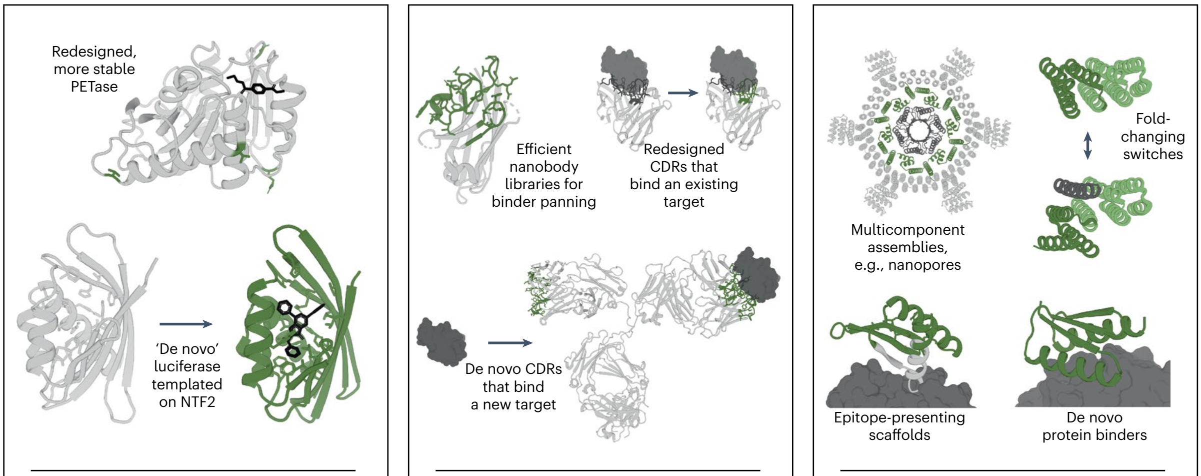
Protein design applications

Machine learning methods have been applied to create new functional designs for a wide range of protein families. This section first focuses on their applications for the design of two protein types with high practical importance: enzymes and antibodies. We then cover other applications, such as improving therapeutic properties and designing protein machines (Fig. 2). With each example, we concisely describe the machine learning model applied, focusing particularly on the nature of the training data (sequence, structure and functional labels) and the training paradigm used (generative or discriminative, supervised or unsupervised) (Fig. 3).

Applications to enzyme design

Improving thermostability

Increasing stability promotes other goals such as improving yield, preventing inactivity and toxicity due to aggregation and operating enzymes at optimal but challenging temperature and solvent conditions (for example, denaturation due to low pH). Enhancing the stability of constructs entering directed evolution may also improve the chances of success by supporting otherwise unstable variants with the desired target function[53–55]. Conventional non-machine learning-based approaches typically find stabilizing mutations by assaying libraries of single substitutions[56,57] or chimeric combinations of natural protein fragments[26,58], often iterating via directed evolution[59,60]. When mutant combinations such as chimeras are assayed, sequence–label models can be used to identify the key mutations leading to increased stability across multiple designs[26,61]. Sequence-based and sequence–structure



	Enzymes	Antibodies	Others
Redesign to enhance existing function	• Enhancing enzyme activity • Improving stability at high temperature or pH	• Improving binding affinity to target • Maintaining affinity while optimizing viscosity, solubility and immunogenicity	• Avoiding human immune reaction • Avoiding chemical modification or protease digestion
Redesign for a new function	• Altering substrate or product specificity	• Decreasing antibody polyreactivity	• Designing allosteric and switchable function
De novo design	• Designing a new scaffold, given a catalytic motif	• Designing new antibody CDRs from scratch, given a target	• Designing a de novo binder, given a target • Scaffold/adjoin functional elements

Fig. 2 | Protein design applications. Examples of machine learning-driven protein design applications for different protein types across the three protein design objective categories. NTF2, nuclear transport factor 2.

models have been surprisingly successful at designing proteins with enhanced stability without the need for labels[1–4,22] to circumvent costly preliminary assays. Although these models do not explicitly predict folding ΔG or melting temperature, the sequence or structure likelihoods output by these generative models trained on natural proteins correlate strongly with stability[12–14,62], enabling success even in small-scale experiments. For example, sequence–structure models were used to design polyethylene terephthalate hydrolase (PETase) variants (one to four mutations)[1] and myoglobin variants (mutating ~50% of the sequence, avoiding 17 amino acids at the heme-binding site)[4] with enhanced stability, assaying only ~20 designs for each. The key challenge is to maintain the original function while making stabilizing mutations, often by avoiding mutations to existing amino acids that are essential for function. This was explored by Sumida et al.[4], who tested 144 redesigned tobacco etch virus (TEV) protease sequences and found that most designs conserving over 50% of the original sequence sustained or enhanced activity, whereas designs with less than 50% retention exhibited loss of activity. Looking forward, new high-throughput stability assays such as those by the Rocklin group (about 800,00 high-quality measurements for over 450 protein domains)[63] may provide the basis to train supervised models[27,64] for thermostability design that generalizes across protein families, with the aim to exceed fully unsupervised methods while still circumventing project-specific data generation.

Altering specificity or activity

Inspired by the fact that natural proteins often evolve multiple unique functions originating from the same family and fold[65–67] or are promiscuous to alternative substrates or reactions[68], a common design strategy is to alter existing proteins. Enzymes can even be modified to catalyze reactions not yet found in nature[42,69,70]. Design of enzymes with the new chemistries has long relied on directed evolution[69,70] or approaches replacing enzyme active sites by structure comparison[42,43]. To reduce the number of mutation–selection rounds required, recent directed evolution efforts have replaced random mutagenesis at each step with library design by sequence–label models updated at each round with the new data[29,30,33,35,71,72].

In a standout example, Schmitt et al. trained a sequence–label conditional variational autoencoder (VAE) model on extensive directed evolution data: 89 Cre-recombinase libraries evolved against different DNA targets, amounting to >2 million protein–DNA sequence pairs[29]. The resulting model was then able to generate new protein sequences conditioned on custom-input DNA sequence, thereby circumventing the need for additional directed evolution campaigns tailored to the new targets. It was used in particular to design a single protein sequence for each of ten new DNA targets not cleaved by the original libraries. Remarkably, four of ten designs showed successful excision at these new targets.

Design a new scaffold, given a functional motif

The ability to embed small functional motifs in new scaffolds enables a more modular design of functional components (for example, to fuse multiple properties). One approach is to combine a compatible structure with a structural arrangement of residues that enable binding or reaction. For example, existing protein domains that already bind or react with some molecule have had active sites replaced to perform new reactions[42,43,73]. Proteins with entirely unrelated function have also been functionalized by this approach. For example, Holst et al. converted an armadillo repeat protein into a new polycarbonate hydrolase by just four mutations[41]. Likewise, force-field methods have been used to design a number of new structures that contain protein-binding motifs, epitopes and fluorophore sites[49,74,75]. To find plausible 3D folds among the huge space of possibilities, conventional de novo methods narrow the space to well-defined topologies with favorable folding interactions. The scope of these approaches may be poised to expand thanks to new methods that propose molecular interactions

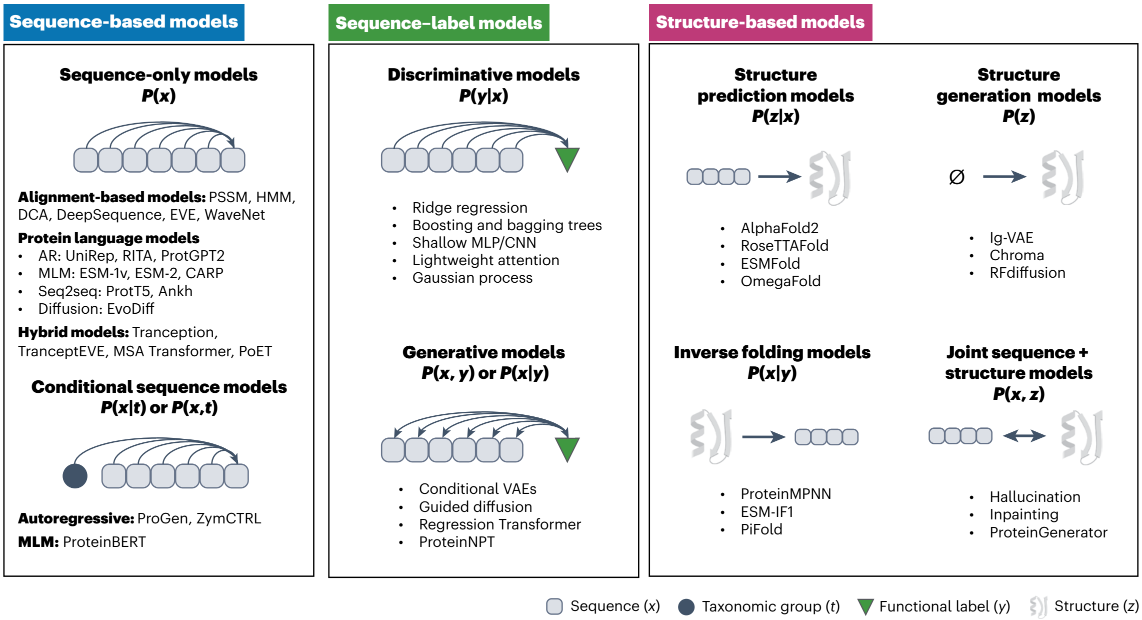


Fig. 3 | Typology of protein design models. The majority of machine learning methods for protein design can be broadly categorized into three groups depending on the data modalities used to train them and the underlying distributions that these approaches seek to model for predictive and generative purposes. AR, autoregressive; MLM, masked language model; MLP, multilayer perceptron.

and identify compatible scaffolds (RifGen, RifDock)[76]. Additionally, scaffold compatibility may be improved by using sequence–structure transformer models to generate modified structures. These methods have already been applied to create new scaffolds for a luciferase functional site by templating on an existing small-molecule-binding family[44] and to customize de novo folds to accommodate functional motifs[23,45,46,48,77–79].

Applications to antibody design

The role of machine learning in antibody design

Antibodies are omnipresent in biomedicine owing to their remarkable specificity and affinity for biomolecules. Previously, obtaining an antibody specific to a therapeutic or scientific target of interest required animal inoculation to exploit the natural process of affinity maturation by somatic hypermutation[80,81]. More recently, antibody-discovery campaigns have shifted toward large-affinity screens such as yeast[82–84] and phage display[85,86]. Yet, these campaigns remain costly and with a low guarantee of success. Machine learning-driven antibody design promises to decrease these costs and increase success rate. While we would like to design a specific antibody directly, computational approaches are currently limited to accelerating certain well-defined steps along the end-to-end discovery process, such as improving the likelihood of successful clones, reducing the need for rounds of affinity maturation, optimization of specificity, reducing polyreactivity or determining the 3D structure of the antibody–target complex.

Enhance features of existing antibodies

The majority of machine learning models used to improve existing antibodies rely on deep sequencing from rounds of selection or deep mutational scans for training data. For instance, Parkinson et al. developed a sequence–label model to predict yeast display affinities on the basis of sequence embeddings pretrained on native antibody sequences. The model was then used to optimize atezolizumab for binding to programmed cell death ligand 1 (PD-L1)[87]. Similarly, Saka et al. trained a sequence-only long short-term memory (LSTM) model on sequences from phage-display selection to optimize an antibody specific to kynurenine[88]. In another instance, Mason et al. developed a discriminative sequence–label convolutional neural network (CNN) by training on deep mutational scan labels. It simultaneously lead to marginal improvement of trastuzumab's affinity for HER2 while also considering viscosity, solubility and immunogenicity[89]. In an application that supports cross-target binding, Liu et al. trained multiple discriminative sequence–label CNN-based models on phage-display data for several binding targets, each target being a separate antibody. The resulting models were used in combination to predict sequences cross-reactive to all targets by binding the Fc[40]. Lastly, Makowski et al. improved binding to the kinase c-MET by training a discriminative sequence–label model on a mutational scan of emibetuzumab[90]. Despite these successes, an important shortcoming of these various methods is that they are usually antibody specific. Their application to other antibodies would require the collection of new, project-specific experimental data, limiting their broader utility.

Three notable strategies have been able to generalize to new antibodies and antigens without project-specific sequencing data. First, Harvey et al. developed a sequence–label model to reduce polyspecificity in antibody and nanobody complementarity-determining regions (CDRs)[32]. The model was trained on deep sequencing of selection experiments for high and low polyreactivity in a naive nanobody library. More than 85% of predicted mutations reduced polyreactivity in three nanobodies. For an anti-angiotensin II type 1 receptor (AT1R) antibody, 3 of 13 predicted mutations decreased polyreactivity by up to 80% while retaining AT1R binding. This suggests that it is possible to reduce the polyspecificity of nanobodies while retaining on-target binding, without antigen-specific data. Second, Hie et al. leveraged a sequence-only protein language model, ESM-1v, to introduce mutations that improve the affinities of seven known antibodies[39]. Here, rather than being trained on antigen selection data specific to the problem at hand, it was instead trained on a large set of protein sequences from diverse families. It was then used to recommend up to 14 single mutations to be screened for binding affinity. Mutants that increased affinity were combined in a second round, resulting in an affinity increase up to 160-fold over that of unmatured antibodies. This implies that generalizable machine learning models can help reduce library size and the number of rounds needed to optimize existing antibodies without modeling the antigen or training on antigen-specific sequencing data. Third, sequence–structure models were used to recommend mutations to an existing antibody sequence, given the 3D structure of the bound antibody–target complex as input[91,92]. In one example, the ESM-IF1 model[62] was used to redesign Ly-1404 and SA58, antibodies for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[91]. With testing of only ~30 designs for each starting point, the top designs had seven and two mutations, and affinity enhanced 26-fold and 11-fold, respectively. The model also showed standout ability to predict the affinity of combinatorial variants of antibodies specific for influenza hemagglutinin. While this inverse folding strategy is restricted to the redesign of antibodies with sufficiently high affinity to obtain 3D structures in complex with target, it may enable de novo antibody design in the future if paired with powerful predictive models proposing new antibody–antigen 3D complexes.

Accelerating affinity campaigns by using smart libraries

One strategy to increase the odds of finding successful antibodies is to design intelligent starting libraries that are enriched for functional antibodies. Although billions of known antibody sequences are currently available (for example, the Observed Antibody Space database[93]), they only represent a small fraction of the pan-human repertoire and we expect to see this number substantially increase as new sequencing methods become available. Furthermore, because the cost of synthesis lags far behind the cost of sequencing, the potential to design custom libraries is still constrained. Consequently, various computational strategies have been developed to approximate the functional antibody sequence space. A simple approach is to generate variations that mimic residue preferences in the CDRs of single-domain antibodies from solved 3D structures. This method has successfully discovered specific nanobodies for two distinct human G protein-coupled receptors[84] and the SARS-CoV-2 receptor-binding domain[83]. An alternative strategy is to design smaller libraries that can be precisely synthesized. For instance, SeqDesign[82], a sequence-only CNN model trained on 12 million publicly available nanobody sequences, generated a diverse library of approximately 185,000 sequences with unique CDR3 regions and both CDR1 and CDR2 identical to those of the germline. This library led to improved expression in yeast, and, despite being 1,000× smaller than typical yeast nanobody campaigns[81], contained low- to moderate-affinity antibodies specific to human serum albumin[84] and green fluorescent protein[83]. Looking ahead, recently proposed computational strategies such as variational synthesis[94] may be used to optimize larger library design.

Design a monobody from scratch, given a target

The closest example to date of machine learning-driven de novo antibody design is the engineering of a single fibronectin 3 (Fn3) domain (in imitation of the immunoglobulin G fold) targeting a conserved epitope on α-elapitoxin[95]. First, 1.6 million Fn3 domains were generated via molecular dynamics simulations on 442 known natural Fn3 domain structures and used to train a structure generation model (similar to Ig-VAE[96]) to then efficiently sample new conformations. Second, the sampled synthetic structures were docked onto the target using a statistical potential based on residue interactions in the PDB. Lastly, specific sequences were generated using a sequence–structure CNN model and further optimized with Rosetta. Roughly 6,000 sequences generated by the method were experimentally screened for binding to five toxins, each of which conserved the target epitope, resulting in one design found to bind three of five toxins. Although the affinity was not reported and a random or target-independent library was not screened for comparison, the corresponding hit rate is much higher than for naive antibody repertoires or random yeast-display libraries, which typically screen several millions of unique sequences to find hits.

For de novo antibody design to be practical, it must substantially decrease the effort needed for screening by reducing necessary throughput, enabling parallelization, decreasing selection rounds or increasing successful design rates. For instance, Shanehsazzadeh et al.[97] generated new heavy-chain CDRs with a model conditioned on an existing antibody–antigen structure (trastuzumab–human HER2) and using framework and light-chain CDR sequences from a known antibody (trastuzumab). The approach led to binding rates higher than those generated by randomly sampling heavy-chain CDRs from existing antibody repertoires. However, the greater practical challenge will be to demonstrate successful designs in cases in which an effective antibody is not already known and therefore the antibody-binding interface, the framework and any CDR sequences are not known. Given the recent success in de novo design of protein folds that bind a target[23,98], a breakthrough in de novo antibody design could be near. This might necessitate innovations such as antibody-specific training, simultaneous sequence and structure optimization[99,100] or more detailed structure modeling like atomistic modeling[101,102].

Other applications

Avoiding human immune reaction

When proteins are injected as therapeutics, they are often thwarted by anti-drug antibodies (ADAs) that can neutralize therapeutic activity and even lead to toxicity[103]. To be safe and effective, a protein therapeutic must avoid binding by existing ADAs at antibody-binding sites (also called B cell epitopes) and avoid eliciting new ADAs by T cell epitopes.

Existing antibodies can be avoided by eliminating B cell epitopes on the protein surface. Strategies include coating the surface with glycans or polymers or ‘resurfacing’ the protein by mutation so that the existing antibodies are no longer compatible[104,105]. For example, Bootwala et al.[106] leveraged an inverse folding model[107] to generate sequences based on L-asparaginase while mutating only the protein surface. The standout design introduced 40 mutations and reduced human ADA binding by 50%, although at the expense of a 50% decrease in both expression and enzymatic activity. With knowledge of antibody-binding hotspots or with approximate antibody-binding probabilities[108], it may be possible to use hotspot-focused resurfacing to achieve similar ADA-binding reduction with fewer mutations and mitigate chances of reduced function.

Newly elicited antibodies can be reduced by eliminating T cell epitopes throughout the linear sequence of a protein by removing the major histocompatibility complex II (MHC-II) display of T cell epitopes or by suppressing T cell receptor complementation. Thus far, protein design to reduce T cell immunogenicity has focused on eliminating MHC-II display[104–108]. This has been partly due to strong capabilities in measuring and predicting MHC-II display[109–113], while the reliability of models to predict T cell receptor complementation remains uncertain[114]. MHC-II-binding sequence–label models have been paired with sequence-only models to identify mutations predicted to both remove epitopes and retain function[8]. Unlike prior epitope-removal approaches that caused function loss[5–7], these approaches left function intact[74]. Similar efforts geared toward MHC-I receptor display, to avoid the killer T cell response, will be important for proteins introduced by gene delivery.

Design a binder, given a target

A major goal of protein design, with massive foreseeable impact in science and medicine, is to produce protein binders for any target, without the need for intensive panning experiments or for an existing known binder. While there have been breakthroughs in force-field design given a target alone, these have typically required thousands of constructs to be assayed[76]. In contrast, recent progress in machine learning approaches has enabled instances of protein binder design requiring fewer than 100 constructs to be assayed[23]. For instance, in a recent two-stage approach, a structure generation model, such as RFdiffusion[22], produces 3D backbones complementary to the structures of protein targets, which are input to an inverse folding model, such as ProteinMPNN[22,74,110], to generate sequences. While this two-step process consisting of structure sampling and sequence design is similar to protein design with Rosetta, it has been able to generate greater breadth of structures with much fewer iterations. When applied to design binders for four protein targets, the approach showed an overall success rate of 18%, eclipsing the performance of prior target-only de novo binder design[22,76,115], potentially due to better shape complementarity of backbones generated by RFdiffusion, in contrast to the formulaic topologies used by non-machine learning methods[76]. Recently, these methods enabled a similar breakthrough in the design of DNA-specific binding proteins[116]. Although there are limitations to current diffusion methods, we expect the increased ability to design customized protein folds, incorporating interfaces and functional motifs, will substantially enhance the practicality of protein design.

Vaccines, machines and more

The aforementioned examples represent some of the most common protein design applications, but there are many more: vaccines protective against future virus evolution[117], vaccine scaffolds that efficiently present immunogens to immune cells[118–122], anti-viral drugs designed to bind and inhibit viral proteins[123–127], switches and sensors that activate a biological pathway in response to target molecules[128–133], axles that perform nanoscale rotation[134] and nanopores for DNA and protein sequencing or detection[135–139].

Presently, innovation in protein design relies on both the existing functions of natural proteins and the customizability of de novo design. Taking nanopore sequencing for example, one can rationally specify dimensions of the pore[47,137,140] to achieve electrical resistance unique to each nucleotide or amino acid that passes through, but strategies to thread proteins through pores still take advantage of existing unfolding machinery from nature[136]. Advances in design methods that further blur the lines between 'find and modify a component to do X' and 'create a component to do X' are the inflection points to watch out for as this field evolves.

Future directions

Data and model scaling

Preliminary analyses of the scaling laws of protein language models[141] have highlighted the benefits of scaling model size and, perhaps, hinted at some of its limits[51]. If trends observed in natural language processing[142] hold for proteins, leveraging larger datasets of protein sequences during training will yield improved generative models and, in turn, increase design quality. Similarly, the performance of protein models, whether for structure prediction[143] or fitness prediction[144], explicitly leveraging homology[145] or with single-sequence input[15,146], increases with a larger number of homologs. Further progress may thus be achieved for specific design efforts by enriching training sets with tailored sequence libraries.

Finer-grained control of design

The ability to design proteins for precise functions and conditions may be facilitated by the growing amount of data quantifying the functions of both known and synthetic proteins. This corpus of data includes resources that compile enzyme classification[147], substrates and reactions[148,149], gene ontologies[150], biophysical properties[63,151], millions of measured effects of mutations[152] and sequences produced by directed evolution[29]. Explicitly modeling this information provides finer control over generated sequences, for instance, by conditioning on broad taxonomic labels[20,153] and, increasingly, on more granular property values[154,155]. Integration with natural language models, as recently introduced for chemical design[156] and proteins[157,158], may eventually offer a more intuitive interface for design.

In silico evaluation

Sequence-only models for design have succeeded in designing functional proteins distant in sequence space from existing proteins[2,3,19–21,36–38]. A direct comparison of the design abilities of different model architectures has nonetheless been difficult due to their application to substantially different protein families. While primarily focused on low mutational depth, benchmarks based on large collections of mutational scan assays such as ProteinGym[18] or FLIP[159] enable thorough baseline comparisons in various design settings and will help guide future model-development efforts. However, an increased accuracy in mutation effect prediction may not necessarily translate into increased design accuracy—that is, a superior ability to generate sequences with the desired functions. While various in silico metrics have emerged as means of comparison between structure-based models, such as recovery of native sequences from existing structures[22,107] or recovery of target structure from designed sequences by AlphaFold prediction[22,23,45,48,115], these are not unbiased measures of protein design success. Regarding sequence recall, a model with enough capacity could easily memorize training sequences and, if prompted appropriately, may perform extremely well on that metric without necessarily being able to extrapolate to new sequences. Regarding structure recall, AlphaFold may serve as a coarse proxy[115], but it is known to fail at discerning stability at the level of individual mutations[160,161]. Although AlphaFold can successfully predict the fold of many de novo proteins designed by methods built on natural protein information, it may fail to generalize to designs that lack sequence and structure features known to nature. More comprehensive evaluation frameworks are emerging[162] and may help to further optimize designs in silico and focus in vitro efforts on higher-quality candidates.

Experimental evaluation

The field still lacks unbiased experimental benchmarks that capture practical design tasks, such as 'given a protein or small-molecule target, design a binder' or 'given a reaction, design an enzyme'. To avoid conflating 'natural protein recall' with design capability, the chosen targets must be 'new to nature', searching for binding or reactions not known in natural examples. The Institute for Protein Design has created some initial benchmark tasks[22,23,48,76] for target protein binding and functional motif scaffolding, although these have yet to address natural protein recall. As most real-life design use cases involve the simultaneous optimization of multiple properties[163], experimental benchmarks should also probe this capability, for instance, via tasks such as 'given a protein, increase its stability while retaining activity' or 'given a protein, alter its function while maintaining stability'.

Toward a unified approach to protein design

Boundaries between the diverse model categories outlined in Box 2 and Fig. 3 are becoming increasingly blurred. Recent developments include models combining structure-aware models together with powerful sequence-based models[164] as well as protein language models trained with a structure-aware vocabulary[165]. Approaches adapted from the natural language processing and computer vision literature[166] are providing effective ways to combine diverse modalities in the same model to learn richer protein representations[167]. Extending machine learning models with biophysics principles may provide the necessary inductive biases to extrapolate beyond the training data[168]. Improved sampling approaches[169–171] will help increase the quality of generated sequences for subsequent experimental validation. The emergence of self-driving laboratories[172,173] that leverage a robust integration of uncertainty-aware models and experimental processes holds the promise of accelerating comprehensive end-to-end design cycles. In the foreseeable future, we anticipate the rise of unified design models, which fuse aspects of sequence-based, sequence–label and structure-based models[64]. This will lead to models capable of supporting both tightly controlled sampling for efficient design iterations, optimizing multiple objectives within specific experimental or cellular contexts, and more challenging de novo tasks aimed at designing proteins with functions beyond those naturally occurring.

References

1. Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
2. Giessel, A. et al. Therapeutic enzyme engineering using a generative neural network. *Sci. Rep.* **12**, 1536 (2022).
3. Fram, B. et al. Simultaneous enhancement of multiple functional properties using evolution-informed protein design. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.09.539914 (2023).
4. Sumida, K. H. et al. Improving protein expression, stability, and function with ProteinMPNN. *J. Am. Chem. Soc.* **146**, 2054–2061 (2024).
5. Schubert, B. et al. Population-specific design of de-immunized protein biotherapeutics. *PLoS Comput. Biol.* **14**, e1005983 (2018).
6. Salvat, R. S. et al. Computationally optimized deimmunization libraries yield highly mutated enzymes with low immunogenicity and enhanced activity. *Proc. Natl Acad. Sci. USA* **114**, E5085–E5093 (2017).
7. Jankowski, W. et al. Mitigation of T-cell dependent immunogenicity by reengineering factor VIIa analogue. *Blood Adv.* **3**, 2668–2678 (2019).
8. Mufarrege, E. F. et al. De-immunized and functional therapeutic (DeFT) versions of a long lasting recombinant α interferon for antiviral therapy. *Clin. Immunol.* **176**, 31–41 (2017).
9. Winterling, K. et al. Development of a novel fully functional coagulation factor VIII with reduced immunogenicity utilizing an in silico prediction and deimmunization approach. *J. Thromb. Haemost.* **19**, 2161–2170 (2021).
10. Zhao, H. et al. Globally deimmunized lysostaphin evades human immune surveillance and enables highly efficacious repeat dosing. *Sci. Adv.* **6**, eabb9011 (2020).
11. Zhao, H. et al. Depletion of T cell epitopes in lysostaphin mitigates anti-drug antibody response and enhances antibacterial efficacy in vivo. *Chem. Biol.* **22**, 629–639 (2015).
12. Hopf, T. A. et al. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
13. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
14. Shin, J.-E. et al. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.* **12**, 2403 (2021).
15. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* **34**, 29287–29303 (2021).
16. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
17. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
18. Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS)* Vol. 36 (2023).
19. Russ, W. P. et al. An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
20. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
21. Lian, X. et al. Deep learning-enabled design of synthetic orthologs of a signaling protein. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.21.521443 (2022).
22. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
23. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
24. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-*N* protein engineering with data-efficient deep learning. *Nat. Methods* **18**, 389–396 (2021).
25. Eid, F.-E. et al. Systematic multi-trait AAV capsid engineering for efficient gene delivery. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.22.521680 (2022).
26. Li, Y. et al. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **25**, 1051–1056 (2007).
27. Pak, M. A., Dovidchenko, N. V., Sharma, S. M. & Ivankov, D. N. New mega dataset combined with deep neural network makes a progress in predicting impact of mutation on protein stability. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.31.522396 (2023).
28. Umerenkov, D. et al. PROSTATA: protein stability assessment using transformers. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.25.521875 (2022).
29. Schmitt, L. T., Paszkowski-Rogacz, M., Jug, F. & Buchholz, F. Prediction of designer-recombinases for DNA editing with generative deep learning. *Nat. Commun.* **13**, 7966 (2022).
30. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl Acad. Sci. USA* **116**, 8852–8858 (2019).
31. Malbranke, C. et al. Computational design of novel Cas9 PAM-interacting domains using evolution-based modelling and structural quality assessment. *PLoS Comput. Biol.* **19**, e1011621 (2023).
32. Harvey, E. P. et al. An in silico method to assess antibody fragment polyreactivity. *Nat. Commun.* **13**, 7554 (2022).
33. Fox, R. J. et al. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).
34. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl Acad. Sci. USA* **110**, E193–E201 (2013).
35. Saito, Y. et al. Machine-learning-guided library design cycle for directed evolution of enzymes: the effects of training data composition on sequence space exploration. *ACS Catal.* **11**, 14615–14624 (2021).
36. Repecka, D. et al. Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
37. Sinai, S., Jain, N., Church, G. M. & Kelsic, E. D. Generative AAV capsid diversification by latent interpolation. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.16.440236 (2021).
38. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
39. Hie, B. L. et al. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01763-2 (2023).
40. Liu, G. et al. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics* **36**, 2126–2133 (2020).
41. Holst, L. H. et al. De novo design of a polycarbonate hydrolase. *Protein Eng. Des. Sel.* **36**, gzad022 (2023).
42. Siegel, J. B. et al. Computational design of an enzyme catalyst for a stereoselective bimolecular Diels–Alder reaction. *Science* **329**, 309–313 (2010).
43. Jiang, L. et al. De novo computational design of retro-aldol enzymes. *Science* **319**, 1387–1391 (2008).
44. Yeh, A. H.-W. et al. De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
45. Anishchenko, I. et al. De novo protein design by deep network hallucination. *Nature* **600**, 547–552 (2021).

46. Verkuil, R. et al. Language models generalize beyond natural proteins. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.21.521521 (2022).
47. Lutz, I. D. et al. Top–down design of protein architectures with reinforcement learning. *Science* **380**, 266–273 (2023).
48. Wang, J. et al. Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
49. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. *Nature* **561**, 485–491 (2018).
50. 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).
51. Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978 (2023).
52. Ferruz, N., Schmidt, S. & Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **13**, 4348 (2022).
53. Bloom, J. D., Wilke, C. O., Arnold, F. H. & Adami, C. Stability and the evolvability of function in a model protein. *Biophys. J.* **86**, 2758–2764 (2004).
54. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl Acad. Sci. USA* **103**, 5869–5874 (2006).
55. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput. Biol.* **4**, e1000002 (2008).
56. Nakatani, K. et al. Increase in the thermostability of *Bacillus* sp. strain TAR-1 xylanase using a site saturation mutagenesis library. *Biosci. Biotechnol. Biochem.* **82**, 1715–1723 (2018).
57. Katano, Y. et al. Generation of thermostable Moloney murine leukemia virus reverse transcriptase variants using site saturation mutagenesis library and cell-free protein expression system. *Biosci. Biotechnol. Biochem.* **81**, 2339–2345 (2017).
58. Richardson, T. H. et al. A novel, high performance enzyme for starch liquefaction. *J. Biol. Chem.* **277**, 26501–26507 (2002).
59. Giver, L., Gershenson, A., Freskgard, P.-O. & Arnold, F. H. Directed evolution of a thermostable esterase. *Proc. Natl Acad. Sci. USA* **95**, 12809–12813 (1998).
60. Bell, E. L. et al. Directed evolution of an efficient and thermostable PET depolymerase. *Nat. Catal.* **5**, 673–681 (2022).
61. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
62. 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.) 8946–8970 (PMLR, 2022).
63. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and protein design. *Nature* **620**, 434–444 (2023).
64. Dieckhaus, H., Brocidiacono, M., Randolph, N. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl Acad. Sci USA* **121**, e2314853121 (2024).
65. Nagano, N., Orengo, C. A. & Thornton, J. M. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765 (2002).
66. Isin, E. M. & Guengerich, F. P. Complex reactions catalyzed by cytochrome P450 enzymes. *Biochim. Biophys. Acta* **1770**, 314–329 (2007).
67. Guengerich, F. P. & Munro, A. W. Unusual cytochrome P450 enzymes and reactions. *J. Biol. Chem.* **288**, 17065–17073 (2013).
68. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
69. Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angew. Chem. Int. Ed. Engl.* **57**, 4143–4148 (2018).
70. Yang, Y. & Arnold, F. H. Navigating the unnatural reaction space: directed evolution of heme proteins for selective carbene and nitrene transfer. *Acc. Chem. Res.* **54**, 1209–1225 (2021).
71. Bedbrook, C. N. et al. Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods* **16**, 1176–1184 (2019).
72. Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol.* **69**, 11–18 (2021).
73. Röthlisberger, D. et al. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
74. Sesterhenn, F. et al. De novo protein design enables the precise induction of RSV-neutralizing antibodies. *Science* **368**, eaay5051 (2020).
75. Yang, C. et al. Bottom–up de novo design of functional proteins with complex structural features. *Nat. Chem. Biol.* **17**, 492–500 (2021).
76. Cao, L. et al. Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).
77. Ingraham, J. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
78. Trippe, B. L. et al. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. In *International Conference on Learning Representations* Vol. 11 (ICLR, 2023).
79. Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nat. Comput. Sci.* **3**, 382–392 (2023).
80. Rajewsky, K. Clonal selection and learning in the antibody system. *Nature* **381**, 751–758 (1996).
81. Teng, G. & Papavasiliou, F. N. Immunoglobulin somatic hypermutation. *Annu. Rev. Genet.* **41**, 107–120 (2007).
82. Boder, E. T., Raeeszadeh-Sarmazdeh, M. & Price, J. V. Engineering antibodies by yeast display. *Arch. Biochem. Biophys.* **526**, 99–106 (2012).
83. Wellner, A. et al. Rapid generation of potent antibodies by autonomous hypermutation in yeast. *Nat. Chem. Biol.* **17**, 1057–1064 (2021).
84. McMahon, C. et al. Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat. Struct. Mol. Biol.* **25**, 289–296 (2018).
85. Almagro, J. C., Pedraza-Escalona, M., Arrieta, H. I. & Pérez-Tapia, S. M. Phage display libraries for antibody therapeutic discovery and development. *Antibodies* **8**, 44 (2019).
86. Ledsgaard, L. et al. Advances in antibody phage display technology. *Drug Discov. Today* **27**, 2151–2169 (2022).
87. Parkinson, J., Hard, R. & Wang, W. The RESP AI model accelerates the identification of tight-binding antibodies. *Nat. Commun.* **14**, 454 (2023).
88. Saka, K. et al. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* **11**, 5852 (2021).
89. Mason, D. M. et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **5**, 600–612 (2021).
90. Makowski, E. K. et al. Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat. Commun.* **13**, 3788 (2022).
91. Shanker, V. R., Bruun, T. U. J., Hie, B. L. & Kim, P. S. Inverse folding of protein complexes with a structure-informed language model enables unsupervised antibody evolution. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.19.572475 (2023).

92. Shanehsazzadeh, A. et al. In vitro validated antibody design against multiple therapeutic antigens using generative inverse folding. In *Generative AI and Biology (GenBio) Workshop, NeurIPS* (2023).

93. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: a diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci.* **31**, 141–146 (2022).

94. Weinstein, E. N. et al. Optimal design of stochastic DNA synthesis protocols based on generative sequence models. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics* (eds Camps-Valls, G., Ruiz, F. J. R. & Valera, I.) 7450–7482 (PMLR, 2022).

95. Eguchi, R. R. et al. Deep generative design of epitope-specific binding proteins by latent conformation optimization. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.22.521698 (2022).

96. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.* **18**, e1010271 (2022).

97. Shanehsazzadeh, A. et al. Unlocking de novo antibody design with generative artificial intelligence. Preprint at *bioRxiv* https://doi.org/10.1101/2023.01.08.523187 (2023).

98. Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).

99. Mahajan, S. P., Ruffolo, J. A., Frick, R. & Gray, J. J. Hallucinating structure-conditioned antibody libraries for target-specific binders. *Front. Immunol.* **13**, 999034 (2022).

100. Lisanza, S. L. et al. Joint generation of protein sequence and structure with RoseTTAFold sequence space diffusion. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.08.539766 (2023).

101. Chu, A. E., Cheng, L., El Nesr, G., Xu, M. & Huang, P.-S. An all-atom protein generative model. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.24.542194 (2023).

102. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.09.561603 (2023).

103. Krishna, M. & Nadler, S. G. Immunogenicity to biotherapeutics — the role of anti-drug immune complexes. *Front. Immunol.* **7**, 21 (2016).

104. Chapman, A. M. & McNaughton, B. R. Scratching the surface: resurfacing proteins to endow new properties and function. *Cell Chem. Biol.* **23**, 543–553 (2016).

105. Remmel, J. L. et al. Combinatorial resurfacing of Dengue envelope protein domain III antigens selectively ablates epitopes associated with serotype-specific or infection-enhancing antibody responses. *ACS Comb. Sci.* **22**, 446–456 (2020).

106. Bootwala, A. et al. Protein re-surfacing of *E. coli* L-asparaginase to evade pre-existing anti-drug antibodies and hypersensitivity responses. *Front. Immunol.* **13**, 1016179 (2022).

107. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems* Vol. 32 (2019).

108. Thadani, N. N. et al. Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).

109. Singh, H. & Raghava, G. P. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* **17**, 1236–1237 (2001).

110. Zhang, L. et al. TEPITOPEpan: extending TEPITOPE for peptide binding prediction covering over 700 HLA-DR molecules. *PLoS ONE* **7**, e30483 (2012).

111. Racle, J. et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).

112. Reynisson, B. et al. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J. Proteome Res.* **19**, 2304–2315 (2020).

113. Racle, J. et al. Machine learning predictions of MHC-II specificities reveal alternative binding mode of class II epitopes. *Immunity* **56**, 1359–1375 (2023).

114. Peters, B., Nielsen, M. & Sette, A. T cell epitope predictions. *Annu. Rev. Immunol.* **38**, 123–145 (2020).

115. Bennett, N. et al. Improving de novo protein binder design with deep learning. *Nat. Commun.* **14**, 2625 (2023).

116. Glasscock, C. J. et al. Computational design of sequence-specific DNA-binding proteins. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.20.558720 (2023).

117. Youssef, N. et al. Deep generative models predict SARS-CoV-2 spike infectivity and foreshadow neutralizing antibody escape. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.08.561389 (2023).

118. Walls, A. C. et al. Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines for SARS-CoV-2. *Cell* **183**, 1367–1382 (2020).

119. Brouwer, P. J. M. et al. Two-component spike nanoparticle vaccine protects macaques from SARS-CoV-2 infection. *Cell* **184**, 1188–1200 (2021).

120. Cohen, A. A. et al. Mosaic nanoparticles elicit cross-reactive immune responses to zoonotic coronaviruses in mice. *Science* **371**, 735–741 (2021).

121. Kang, Y.-F. et al. Rapid development of SARS-CoV-2 spike protein receptor-binding domain self-assembled nanoparticle vaccine candidates. *ACS Nano* **15**, 2738–2752 (2021).

122. Nguyen, B. & Tolia, N. H. Protein-based antigen presentation platforms for nanoparticle vaccines. *NPJ Vaccines* **6**, 70 (2021).

123. Karoyan, P. et al. Human ACE2 peptide-mimics block SARS-CoV-2 pulmonary cells infection. *Commun. Biol.* **4**, 197 (2021).

124. Glasgow, A. et al. Engineered ACE2 receptor traps potently neutralize SARS-CoV-2. *Proc. Natl Acad. Sci. USA* **117**, 28046–28055 (2020).

125. Torchia, J. A. et al. Optimized ACE2 decoys neutralize antibody-resistant SARS-CoV-2 variants through functional receptor mimicry and treat infection in vivo. *Sci. Adv.* **8**, eabq6527 (2022).

126. Cao, L. et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* **370**, 426–431 (2020).

127. Hunt, A. C. et al. Multivalent designed proteins neutralize SARS-CoV-2 variants of concern and confer protection against infection in mice. *Sci. Transl. Med.* **14**, eabn1252 (2022).

128. Zhang, J. Z. et al. Thermodynamically coupled biosensors for detecting neutralizing antibodies against SARS-CoV-2 variants. *Nat. Biotechnol.* **40**, 1336–1340 (2022).

129. Leonard, A. C. & Whitehead, T. A. Design and engineering of genetically encoded protein biosensors for small molecules. *Curr. Opin. Biotechnol.* **78**, 102787 (2022).

130. Quijano-Rubio, A. et al. De novo design of modular and tunable protein biosensors. *Nature* **591**, 482–487 (2021).

131. Langan, R. A. et al. De novo design of bioactive protein switches. *Nature* **572**, 205–210 (2019).

132. Ng, A. H. et al. Modular and tunable biological feedback control using a de novo protein switch. *Nature* **572**, 265–269 (2019).

133. Lee, G. R. et al. Small-molecule binding and sensing with a designed protein family. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.01.565201 (2023).

134. Courbet, A. et al. Computational design of mechanically coupled axle-rotor protein assemblies. *Science* **376**, 383–390 (2022).

135. Huang, G., Willems, K., Soskine, M., Wloka, C. & Maglia, G. Electro-osmotic capture and ionic discrimination of peptide and protein biomarkers with FraC nanopores. *Nat. Commun.* **8**, 935 (2017).

136. Zhang, S. et al. Bottom–up fabrication of a proteasome–nanopore that unravels and processes single proteins. *Nat. Chem.* **13**, 1192–1199 (2021).

137. Shimizu, K. et al. De novo design of a nanopore for single-molecule detection that incorporates a β-hairpin peptide. *Nat. Nanotechnol.* **17**, 67–75 (2022).
138. Alfaro, J. A. et al. The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* **18**, 604–617 (2021).
139. Berhanu, S. et al. Sculpting conducting nanopore size and shape through de novo protein design. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.20.572500 (2023).
140. Xu, C. et al. Computational design of transmembrane pores. *Nature* **585**, 129–134 (2020).
141. Hesslow, D., Zanichelli, N., Notin, P., Poli, I. & Marks, D. RITA: a study on scaling up generative protein sequence models. *Workshop on Computational Biology, ICML* (2022).
142. Hoffmann, J. et al. Training compute-optimal large language models. *Adv. Neural Inf. Process. Syst.* **35**, 30016–30030 (2022).
143. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
144. Notin, P. et al. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *Proceedings of the 39th International Conference on Machine Learning* 16990–17017 (PMLR, 2022).
145. Notin, P. et al. TranceptEVE: Combining family-specific and family-agnostic models of protein sequences for improved fitness prediction. *Learning Meaningful Representations of Life Workshop, NeurIPS* (2022).
146. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
147. Kanehisa, M. Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol. Biol.* **1611**, 135–145 (2017).
148. Mendez, D. et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
149. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304–305 (2000).
150. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
151. Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D. & Gromiha, M. M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* **49**, D420–D424 (2021).
152. Rubin, A. F. et al. MaveDB v2: a curated community database with over three million variant effects from multiplexed functional assays. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.29.470445 (2021).
153. Munsamy, G., Lindner, S., Lorenz, P. & Ferruz, N. ZymCTRL: a conditional language model for the controllable generation of artificial enzymes. In *Machine Learning for Structural Biology Workshop*, *NeurIPS* (2022).
154. Born, J. & Manica, M. Regression Transformer: concurrent sequence regression and generation for molecular language modeling. *Nat. Mach. Intell.* **5**, 432–444 (2023).
155. Notin, P., Weitzman, R., Marks, D. S. & Gal, Y. ProteinNPT: improving protein property prediction and design with non-parametric transformers. In *Advances in Neural Information Processing Systems* Vol. 36 (2023).
156. Bran, A. M., Cox, S., White, A. D. & Schwaller, P. ChemCrow: augmenting large-language models with chemistry tools. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2304.05376 (2023).
157. Liu, S. et al. A text-guided protein design framework. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.04611 (2023).
158. Hie, B. et al. A high-level programming language for generative protein design. Preprint at *bioRxiv* https://doi.org/10.1101/2022.12.21.521526 (2022).
159. Dallago, C. et al. FLIP: benchmark tasks in fitness landscape inference for proteins. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2021).
160. Pak, M. A. et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. *PLoS ONE* **18**, e0282689 (2023).
161. AlphaFold Protein Structure Database. Frequently asked questions. *AlphaFold Protein Structure Database* https://alphafold.ebi.ac.uk/faq (2022).
162. Johnson, S. R. et al. Computational scoring and experimental evaluation of enzymes generated by neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.04.531015 (2023).
163. Tagasovska, N. et al. A Pareto-optimal compositional energy-based model for sampling and optimization of protein sequences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.10838 (2022).
164. Zheng, Z. et al. Structure-informed language models are protein designers. In *International Conference on Machine Learning* Vol. 40 (PMLR, 2023).
165. Su, J. et al. SaProt: protein language modeling with structure-aware vocabulary. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.01.560349 (2023).
166. Radford, A. et al. Learning transferable visual models from natural language supervision. *In International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
167. Xu, M., Yuan, X., Miret, S. & Tang, J. ProtST: multi-modality learning of protein sequences and biomedical texts. In *International Conference on Machine Learning* Vol. 40 (PMLR, 2023).
168. Malbranke, C., Bikard, D., Cocco, S., Monasson, R. & Tubiana, J. Machine learning for evolutionary-based and physics-inspired protein design: current and future synergies. *Curr. Opin. Struct. Biol.* **80**, 102571 (2023).
169. Frey, N. C. et al. Protein discovery with discrete walk–jump sampling. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.12360 (2023).
170. Darmawan, J. T., Gal, Y. & Notin, P. Sampling protein language models for functional protein design. In *Generative AI and Biology (GenBio) Workshop, NeurIPS* (2023).
171. Kirjner, A. et al. Optimizing protein fitness using Gibbs sampling with graph-based smoothing. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2307.00494 (2023).
172. Rapp, J. T., Bremer, B. J. & Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **1**, 97–107 (2024).
173. Yu, T., Boob, A. G., Singh, N., Su, Y. & Zhao, H. In vitro continuous protein evolution empowered by machine learning and automation. *Cell Syst.* **14**, 633–644 (2023).
174. Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* **108**, E1293–E1301 (2011).
175. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
176. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.19.492714 (2023).
177. Elnaggar, A. et al. ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
178. Elnaggar, A. et al. Ankh: optimized protein language model unlocks general-purpose modelling. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.06568 (2023).
179. Rao, R. M. et al. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning* 8844–8856 (PMLR, 2021).
180. Truong, T. F. Jr. & Bepler, T. PoET: a generative model of protein families as sequences-of-sequences. *Advances in Neural Information Processing Systems* Vol. 36 (2023).

181. Alamdari, S. et al. Protein generation with evolutionary diffusion: sequence is all you need. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.11.556673 (2023).
182. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).
183. Bryant, D. H. et al. Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
184. Zhu, D. et al. Optimal trade-off control in machine learning-based library design, with application to adeno-associated virus (AAV) for gene therapy. *Sci. Adv.* **10**, eadj3786 (2024).
185. Heinzinger, M. et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).
186. Stärk, H., Dallago, C., Heinzinger, M. & Rost, B. Light attention predicts protein location from the language of life. *Bioinform. Adv.* **1**, vbab035 (2021).
187. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
188. Gruver, N. et al. Protein design with guided discrete diffusion. In *Advances in Neural Information Processing Systems* Vol. 36 (2023).
189. Blaabjerg, L. M. et al. Rapid protein stability prediction using deep learning representations. *eLife* **12**, e82593 (2023).
190. Baek, M. Efficient and accurate prediction of protein structures and interactions using RoseTTAFold. *Acta Crystallogr. A Found. Adv.* **78**, a235 (2022).
191. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.21.500999 (2022).
192. Anand, N., Eguchi, R. & Huang, P.-S. Fully differentiable full-atom protein backbone generation. In *Deep Generative Models for Highly Structured Data Workshop, ICLR* (2019).
193. Wu, K. E. et al. Protein structure generation via folding diffusion. *Nat. Commun.* **15**, 1059 (2024).
194. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L. & Dror, R. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations* Vol. 9 (2021).
195. Gao, Z., Tan, C., Chacón, P. & Li, S. Z. PiFold: toward effective and efficient protein inverse folding. In *International Conference on Learning Representations* Vo. 11 (2023).
196. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems* Vol. 29 (2016).
197. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations* Vol. 5 (2017).
198. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A. & Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Process. Mag.* **34**, 18–42 (2017).
199. Veličković, P. et al. Graph attention networks. In *International Conference on Learning Representations* Vol. 6 (2018).
200. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
201. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
202. Castro, E. et al. Transformer-based protein generation with regularized latent space optimization. *Nat. Mach. Intell.* **4**, 840–851 (2022).
203. Notin, P., Hernández-Lobato, J. M. & Gal, Y. Improving black-box optimization in VAE latent space using decoder uncertainty. *Adv. Neural Inf. Process. Syst.* **34**, 802–814 (2021).

Acknowledgements

We thank members of the Marks lab for valuable discussions. P.N. was supported by GSK, the UK Engineering and Physical Sciences Research Council (EPSRC ICASE award no. 18000077) and a Chan Zuckerberg Initiative Award (Neurodegeneration Challenge Network, CZI2018-191853). Y.G. holds a Turing AI Fellowship (Phase 1) at the Alan Turing Institute, which is supported by EPSRC grant reference V030302/1. C.S. is supported by the National Resource for Network Biology (NRNB, P41GM103504). D.S.M. holds a Ben Barres Early Career Award from the Chan Zuckerberg Initiative as part of the Neurodegeneration Challenge Network (CZI2018-191853) and is supported by a NIH Transformational Research Award (TR01 1R01CA260415).

Competing interests

D.M. is an advisor for Dyno Therapeutics, Octant, Jura Bio, Tectonic Therapeutic and Genentech and a cofounder of Seismic. N.R. is employed by Seismic. C.S. is on the scientific advisory board of CytoReason. The other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Pascal Notin, Nathan Rollins or Debora Marks.

Reprints and permissions information is available at www.nature.com/reprints.

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