Efficient exploration of sequence space by sequence-guided protein engineering and design

*Ben E. Clifton, Dan Kozome, and Paola Laurino**

Protein Engineering and Evolution Unit, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna, Okinawa, Japan 904-0495.

ABSTRACT. The rapid growth of sequence databases over the past two decades means that protein engineers faced with optimizing a protein for any given task will often have immediate access to a vast number of related protein sequences. These sequences encode information about the evolutionary history of the protein and the underlying sequence requirements to produce folded, stable, and functional protein variants. Methods that can take advantage of this information are an increasingly important part of the protein engineering toolkit. In this perspective, we discuss the utility of sequence data in protein engineering and design, focusing on recent advances in three main areas: the use of ancestral sequence reconstruction as an engineering tool to generate thermostable and multifunctional proteins, the use of sequence data to guide engineering of multipoint mutants by structure-based computational protein design, and the use of unlabeled sequence data for unsupervised and semi-supervised machine learning, allowing the generation of diverse and functional protein sequences in unexplored regions of sequence space. Altogether, these methods enable the rapid exploration of sequence space within regions enriched with

functional proteins and therefore have great potential for accelerating the engineering of stable, functional, and diverse proteins for industrial and biomedical applications.

Introduction

Proteins have a broad range of medical, industrial, and scientific applications, but often need to be subjected to protein engineering to alter their function or optimize properties such as thermostability, catalytic activity, solubility or stereoselectivity before these applications can be fully realized. In most cases, this is achieved using rational design or directed evolution, an iterative process of random or semi-random mutagenesis followed by screening or selection to identify protein variants of high fitness (i.e., those that display the properties desired by the protein engineer). Although these strategies can be highly effective, they are generally limited to consideration of sequences very similar to the initial protein sequence[1], because >30% of mutations have a negative impact on protein stability and function[2–4], and simultaneous introduction of multiple random mutations often results in loss of function[2,4,5]. On the other hand, any method for protein engineering or design that attempts to explore regions of sequence space more distant from natural sequences, seeking larger or faster improvements in protein fitness, needs to contend with the vastness and emptiness of sequence space: there are an incomprehensibly large number of possible protein sequences (for example, $\sim 10^{130}$ for a protein 100 amino acids in length), yet a miniscule fraction of them exhibit any given function[6,7]. Optimization of protein fitness by exhaustive exploration of sequence space, either experimentally or computationally, is unfeasible; complete randomization of even 10 amino acids would result in a theoretical library size ($>10^{13}$) unattainable by most methods[8]. Thus, any more extensive exploration of the sequence

space surrounding a natural protein than can be achieved by directed evolution requires us to consider substitutions that can be predicted in advance to be advantageous or at least neutral.

A useful source of information that can be used to guide the exploration of sequence space towards functional proteins is the data readily available from protein sequence databases. The UniProt Knowledgebase currently contains ~219 million non-redundant protein sequences (2021_03 release) and is continuing to experience rapid growth, driven mainly by large-scale eukaryotic genome sequencing projects and improved methods for metagenome assembly[9]. Although most of these sequences are not experimentally characterized, we can generally assume that they have experienced natural selection to maintain some kind of structure and function (ignoring complications such as pseudogenization, intrinsically disordered proteins, and sequencing errors), and can often make a reasonable prediction about their molecular function. It is usually possible to obtain a large and diverse set of homologous sequences (up to ~10^6) for any given protein through a simple search of protein sequence databases using tools like BLAST[10] or HMMER[11]; for example, ~54% of protein families defined by the Pfam database[12] (version 3.40) are represented by more than 500 sequences. These sequences contain valuable information about the evolutionary history of the protein family and the underlying sequence requirements to produce folded, stable, and functional proteins, such as conservation of catalytic residues and interdependencies between residues. Protein engineers have been taking advantage of this information for decades, for example, to guide the selection of consensus mutations that improve protein thermostability[13,14] or mutations that interconvert the catalytic or binding specificities of two homologous proteins[15,16]. More recently, however, new strategies for protein engineering and design have been developed that take greater advantage of the vast quantity of sequence data now

available, allowing us to extract more information about sequence-function relationships from raw sequence data, explore regions of sequence space increasingly distant from natural proteins, and witness larger improvements in protein properties with less experimental effort.

In this perspective, we discuss recent advances in three main categories of protein engineering and design methods where sequence data can be usefully incorporated: (i) methods based on phylogenetic analysis, including consensus design and ancestral sequence reconstruction (ASR), (ii) structure-based computational protein design, and (iii) machine learning. Recent work has established techniques such as ASR and sequence-guided structure-based design as reliable tools for generating thermostable and functionally diverse proteins that can be useful starting points for further engineering, and has demonstrated the enormous potential of machine learning in designing functional and highly diverse proteins from unexplored regions of sequence space. We limit our discussion of each method to approaches that incorporate data from sequence databases, and invite the interested reader to consult other recent reviews for broader discussions of structure-based design[17], machine learning[18,19], and data-driven protein engineering[1,20–22], as well as applications of sequence data in genome mining and enzyme discovery[23].

Phylogenetic methods: consensus design and ancestral sequence reconstruction

Consensus design is perhaps one of the oldest sequence-based approaches for protein engineering[13,14], but remains an effective and straightforward method for engineering protein thermostability[24]. In this method, a multiple sequence alignment of homologous sequences from a particular protein family is constructed, and the most frequent (consensus) amino acid at each position is identified (Fig. 1). Deviations from the consensus sequence in any given protein are

considered to be, on average, destabilizing. Thus, protein variants with improved thermostability can be obtained through amino acid substitutions that restore the consensus amino acid at each position; these substitutions can be introduced individually into a single protein[14], incorporated into libraries for directed evolution[25], or combined to create a full-length consensus sequence[26]. Although there are numerous recent examples where consensus design has been used successfully to increase protein thermostability[27–31], the method suffers from sensitivity to phylogenetic bias caused by inclusion of closely related sequences in the analysis, and does not account for amino acid covariation at different positions in the multiple sequence alignment[24]. Recent work has continued to refine, generalize, and systematically validate the consensus design methodology[31–34]; for example, using a standardized workflow with very large sequence datasets (1,355 to 14,474 sequences), Sternke *et al*. applied consensus design to six structurally and functionally diverse protein families and achieved increased thermostability over naturally occurring homologs (including those from thermophilic organisms) in four out of six cases[31]. Nonetheless, when proteins obtained by consensus design and ASR using the same multiple sequence alignment have been compared side-by-side, proteins obtained by ASR have usually shown higher thermostability[35–37], although there are exceptions[38].

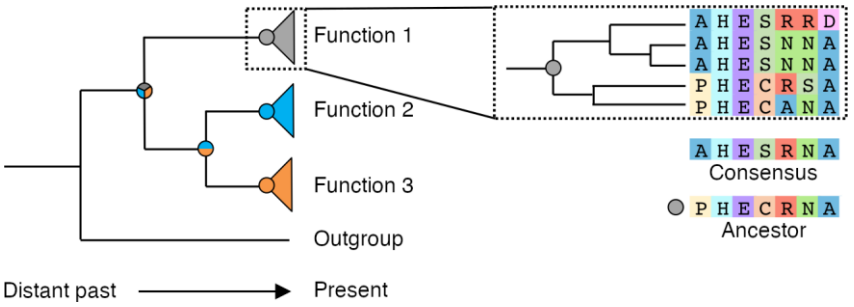


Figure 1. Consensus design and ancestral sequence reconstruction. Both methods use a multiple sequence alignment of sequences homologous to the protein of interest. The consensus

sequence is identified by calculating the most frequent amino acid at each position, while ancestral sequences at each node in a phylogenetic tree are inferred using a statistical phylogenetic analysis based on a multiple sequence alignment and a statistical model of sequence evolution.

Ancestral sequence reconstruction (ASR) is another technique based on multiple sequence alignment of homologous sequences that is becoming widely used for protein engineering, particularly to improve protein thermostability[39,40]. This method uses maximum likelihood or Bayesian methods for statistical phylogenetic analysis to reconstruct plausible sequences of the extinct ancestors of a modern protein family, based on a multiple sequence alignment and phylogenetic tree of the modern proteins and a statistical model of sequence evolution[39] (Fig. 1). The use of ASR in protein engineering was originally motivated by early studies that used this technique to reconstruct ancient proteins from extinct organisms (up to ~3.5 billion years old), in order to experimentally measure their thermostability and thereby infer trends in thermophilicity and environmental temperatures over geological timescales[37,41–44]. These studies showed that reconstructed ancient proteins had consistently higher thermostability (up to 40 °C) than their modern descendants from mesophilic organisms, which was interpreted as evidence that they originated from thermophilic organisms that lived in comparatively hot environments. At the same time, these studies established ASR as a reliable method for generating thermostable proteins based solely on sequence data, suggesting a possible application of this technique in protein engineering.

ASR has now been successfully applied to a variety of protein families to engineer remarkably thermostable biocatalysts[45–52], biopharmaceuticals[53–55] and research tools[56–58], including

carboxylic acid reductases (T_m up to 35 °C higher than characterized extant proteins)[49], amino acid-binding proteins (30 °C)[56], ketol-acid reductoisomerases (30 °C)[46], haloalkane dehalogenases (24 °C)[48,59], and diterpene cyclases (13 °C)[47]. Surprisingly, major improvements in thermostability have been observed even when reconstructing more evolutionarily recent proteins that are not predicted to have originated from thermophilic organisms[46,57,60]; for example, reconstructed cytochrome P450 enzymes and flavin-containing monooxygenases putatively derived from ancestral vertebrates showed T_m values up to 30 °C and 22 °C higher than extant homologs, respectively[46,57]. Although the origin of stabilizing mutations in such cases is not entirely understood, systematic biases in the commonly used maximum likelihood method for ASR may be partly responsible[60–62]; for example, sequence similarity between ancestral and consensus sequences based on the same sequence dataset has suggested a bias of ASR towards the consensus sequence at ambiguously reconstructed positions[25,60], although this bias cannot fully explain the higher stability of ancestral proteins compared with extant proteins[35,36]. There is a need to better understand the source of stabilizing mutations in reconstructed ancestral sequences to predict which protein families will be amenable to ASR as a method to engineer thermostability and to guide the choice of ancestral nodes for experimental characterization; nonetheless, the examples listed above provide empirical evidence that ASR can be used to substantially increase protein thermostability when applied to a dataset of sufficient sequence diversity, even if the resulting ancestral sequences are not evolutionarily ancient (<300 million years old).

When applied to functionally diverse protein families, ASR can also be used to engineer proteins that are more promiscuous or multifunctional than modern proteins[43,47,48,50–52,59,63,64], which may be particularly useful for biocatalysis or to create evolvable starting points for further

engineering[65]. This application of ASR is motivated by the hypothesis that there is a general trend from functional promiscuity to functional specificity in protein families over evolutionary time, based, for example, on the observation that functionally diverse protein families often originate from subfunctionalization of a promiscuous or multifunctional ancestral protein[66]. Although the evidence for this hypothesis is not conclusive[61,67], it is nonetheless true that ASR has been used successfully to engineer proteins with broader specificity that have useful applications. For example, Nakano *et al*. reconstructed an ancestral L-amino acid oxidase with broad specificity for L-amino acids starting from an extant enzyme specific for L-arginine and L-lysine. This enzyme showed oxidase activity on 13 proteinogenic L-amino acids and various non-proteinogenic L-amino acids, enabling production of a wide range of enantiopure D-amino acids[52,68].

The main obstacle to the use of ASR as a protein engineering tool is perhaps the need for extensive manual curation of the sequence dataset and multiple sequence alignment used for phylogenetic tree inference and reconstruction of ancestral sequences. In particular, the quality of the multiple sequence alignment is critical for the accuracy of the reconstruction, and erroneous gaps in the alignment may lead to artefactual insertions in the ancestral proteins[69–71]. Computational tools specifically targeted towards protein engineers may be useful for increasing the accessibility of ASR as a protein engineering tool, in which case the careful treatment of statistical robustness necessary to draw evolutionary conclusions from reconstructed ancestral proteins[72] is not strictly required. For example, the FireProt^{ASR} tool fully automates the ASR workflow, including sequence curation and gap reconstruction, allowing researchers without experience in phylogenetic analysis to use ASR for protein engineering[73]. However, further

improvements in multiple sequence alignment and gap reconstruction algorithms may be needed before automated methods can achieve the same performance as the standard manual approach.

Structure-based design guided by sequence information

Structure-based computational protein design has shown great success in *de novo* design of protein structures, folds, and assemblies unseen in nature[17,74], and has also been used for rational or semi-rational engineering of protein thermostability, catalytic activity, and stereoselectivity[75–78]. However, the structure-based energy calculations used for computational protein design have limited accuracy, because there is an inevitable tradeoff between accuracy and computational feasibility in the energy functions used for these calculations, making it challenging to accurately predict the effects of individual substitutions on protein structure or function for engineering purposes[79,80]. This is a particular problem when attempting to introduce multiple substitutions simultaneously, because a single deleterious substitution can jeopardize the beneficial effects of other substitutions[80]. Several computational tools have been developed that attempt to address these limitations of structure-based design by incorporating information from sequence data into the design algorithm, such that the choice of potential substitutions is restricted or biased towards those that are predicted to be tolerated, based on their occurrence in natural homologs of the target protein[80–88]. A particular advantage of this approach is that substitutions that improve the target property (e.g., thermostability) but decrease protein fitness *via* their effect on another property (e.g., solubility or function) can be avoided, because these substitutions are less likely to be found in natural proteins[89].

This combination of sequence-based and structure-based design has been successfully implemented in two computational tools, PROSS and FireProt, that aim to redesign proteins for increased thermostability and expression[83,84,87,90–92]. PROSS uses the observed amino acid frequencies at each position in a multiple sequence alignment to define a set of "allowed" substitutions in the target protein based on their occurrence in homologous proteins[87] (Fig. 2). Allowed substitutions that have a stabilizing effect on protein structure are predicted using structure-based energy calculations, and mutually compatible combinations of these substitutions are then predicted by combinatorial protein design in Rosetta[93], yielding a small number of designed sequences for experimental testing with substitutions at up to ~10% of positions. Recently, a systematic, community-wide evaluation demonstrated the high success rate of this method for engineering proteins with high thermostability and expression[94]. In this evaluation, 12 independent research groups used PROSS to improve soluble expression of 14 challenging, mostly eukaryotic, proteins in *Escherichia coli*, testing 1–6 designs for each protein. 9/14 proteins showed an increase in soluble expression, 9/10 well-expressed proteins showed an increase in T_m (5.4 °C to 27 °C). 6/7 functionally characterized proteins showed similar function to the wild-type protein for at least one design, indicating that thermostability and activity do not trade off in the designed proteins. The success rate of this method in achieving soluble expression of recalcitrant proteins in *E. coli* is particularly notable given that low thermostability is far from the only cause of poor heterologous expression. In the FireProt method, potentially stabilizing "energy-based" and "evolution-based" substitutions are identified separately by structure-based energy calculations and consensus design, respectively[83,84]. False positives in the energy-based substitutions are then filtered using an evolution-based criterion, and *vice versa*. Finally, compatible substitutions are combined to yield a small number of multipoint variants for experimental testing. In the case of

the haloalkane dehalogenase DhaA, combination of eight energy-based substitutions and three evolution-based substitutions using FireProt yielded a T_m increase of 24.6 °C[83]. The increase in T_m resulted from a near-additive contribution from each set of substitutions (16.2 °C and 9.6 °C, respectively), illustrating the complementarity of these two approaches[95].

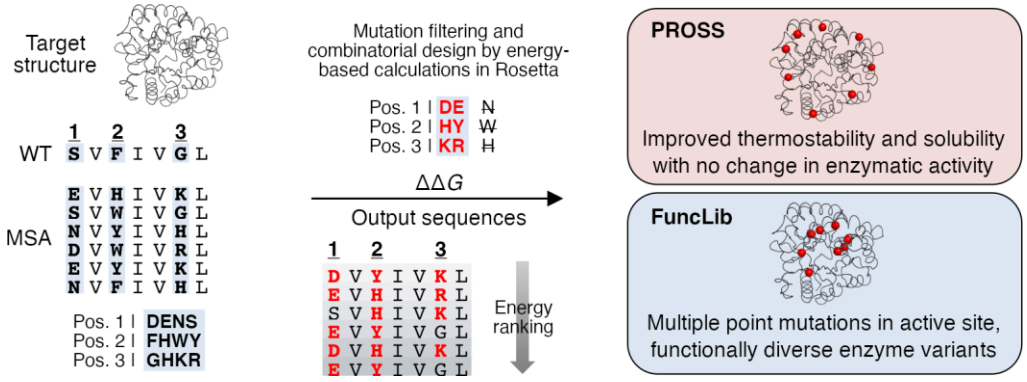


Figure 2. The PROSS and FuncLib methods for incorporating sequence data in structure-based protein design. Evolutionarily "allowed" substitutions at each position (or pre-defined positions in the active site) in the target protein (WT) are identified from a multiple sequence alignment (MSA), reducing the search space for combinatorial design. The allowed substitutions are analyzed further using structure-based energy calculations to predict stabilizing substitutions or eliminate destabilizing substitutions. A small number of sequence designs with multiple point substitutions are then predicted by combinatorial protein design in Rosetta.

A similar approach of using sequence data as a restraint for structure-based engineering of enzyme function is used by the FuncLib method[81]. The aim of this method is to engineer a small number of stable and functionally diverse enzyme variants with multiple substitutions in the active site. Rather than optimizing enzyme activity explicitly (which is more challenging and requires detailed knowledge of enzyme mechanism), the aim is to engineer variants that have a stable, preorganized active site compatible with enzyme activity, which can then be subjected to low-

throughput experimental screening. Similar to the PROSS method, a set of allowed substitutions within the active site is first defined based on evolutionary conservation and structure-based energy calculations (to exclude destabilizing substitutions); multipoint variants containing three to five allowed substitutions are then designed using combinatorial protein design in Rosetta (Fig. 2). In a recent example that takes great advantage of this method, Bengel *et al*. used FuncLib to redesign a promiscuous nicotinamide *N*-methyltransferase to create a panel of enzymes for regioselective *N*-alkylation of pyrazoles[96]. More than 90% of enzyme variants showed activity on at least one pyrazole substrate, and despite the low regioselectivity of the parent enzyme (57 to 67% depending on the substrate), different enzyme variants yielding increased activity (up to 118-fold), regioselectivity (up to >99%), and in some cases regiodivergence could be identified for a diverse range of substrates. Such large improvements in enzyme function are rarely seen in a single round of mutagenesis and screening using conventional strategies, even for a single substrate[97].

Statistical modeling of protein sequence and function: machine learning and other approaches

Machine learning is used for analysis of large and complex datasets in many fields of biology and is beginning to find application in protein engineering[18,98]. In general, machine learning seeks to identify patterns in data without attempting to explicitly model the underlying physical or biological processes that generated the data. Currently, the machine learning methods most widely used in protein engineering are supervised methods, in which a quantitative model of the protein fitness landscape is inferred based on labeled training data (i.e., a set of protein sequences and their measured fitness), and then used to predict the fitness of uncharacterized proteins[18,19]. While this

approach has shown recent success[99–102], it relies on experimental data for training and usually requires multiple iterations of model training and experimental characterization of protein variants[101,103–105]. It would be ideal if we could instead use readily available sequence data to learn the underlying sequence patterns or "design rules" that define a particular protein structure or function, and then generate new and improved sequences that display those patterns, while eliminating or minimizing the need for experimental training data. Recent work has made some exciting progress towards this goal using unsupervised learning (which requires only unlabeled protein sequence data) and semi-supervised learning (which combines supervised and unsupervised learning). Because the topic of machine learning in protein engineering has been covered in recent reviews[1,18,19,106–108], in this section, we focus on several key concepts and experimentally validated advances that are particularly relevant to the applications of sequence data in protein engineering and design.

Whereas *supervised* methods in machine learning-based protein engineering are used to train *discriminative* models, which explicitly model the relationship between sequence and fitness based on a labeled training dataset, *unsupervised* machine learning is used to train *generative* models, which model the probability distribution underlying an unlabeled training set of protein sequences (Fig. 3). Novel sequences that recapitulate the properties of the training sequences can then be obtained by random sampling of sequences from the resulting probability distribution. In other words, generative modeling can be used to learn the region of sequence space associated with a particular function, allowing generation of new sequences from within this space. Various types of deep generative models (i.e., generative models obtained by deep learning) have been applied to protein sequence data, including variational autoencoders[109–112], generative adversarial

networks[113–115], and deep autoregressive models[116,117]. Models from the field of natural language processing, used for tasks such as machine translation, text summarization, and text generation, are frequently used, capitalizing on useful (albeit imperfect) analogies between protein sequences and written text[117–122]. For example, state-of-the-art language models are designed to capture long-range dependencies between tokens (words or characters), which is necessary to learn how the probability of encountering a token at a given position varies depending on the surrounding tokens (e.g., how the meaning of a word changes depending on the context)[123]. This feature may also be useful for capturing long-range dependencies between amino acids within a protein sequence (i.e., covariation between amino acids that are distant in the protein sequence but close in the protein structure)[118].

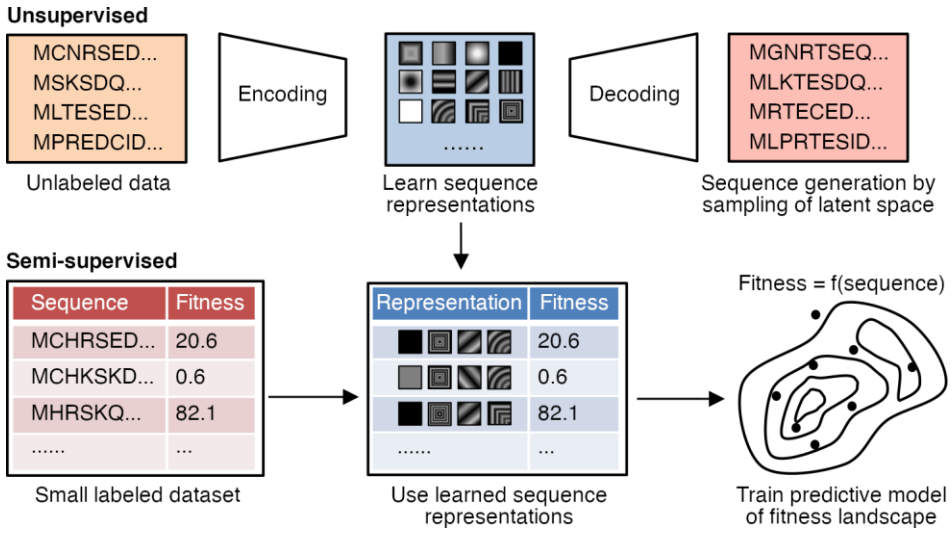


Figure 3. Unsupervised and semi-supervised machine learning approaches for protein engineering and design. Using unsupervised learning, the probability distribution underlying an input sequence dataset can be learned, allowing generation of new sequences by sampling from the probability distribution. Some unsupervised machine learning algorithms, such as variational autoencoders (depicted here), are also capable of learning informative representations of protein

sequence. In semi-supervised methods, these representations can be used to improve the predictive model of the fitness landscape inferred by supervised learning.

Deep generative models have been shown to capture physicochemical, structural, evolutionary, and functional information from protein sequences[117,118] and have shown state-of-the-art performance on problems such as variant effect prediction[111,124], but there are still few examples where proteins have been designed using generative models and experimentally validated by functional characterization[112,114,116,125]. In one recent example, Repecka *et al*. trained a generative adversarial network on 16,706 malate dehydrogenase sequences and used the resulting model to generate synthetic protein sequences[114]. The generated proteins were shown to reproduce various properties of the natural proteins at the sequence level (e.g., sequence variability at each position and conservation of key catalytic residues), while occupying different regions of sequence space and showing higher diversity compared with the natural sequences. A small subset of generated sequences was then experimentally characterized; 13 out of 55 (24%) of the designs were soluble and functional when expressed in *E. coli*, with the most divergent of these proteins having 66% sequence identity (106 substitutions) to a natural sequence. Altogether, these studies demonstrate that deep generative models can generate diverse, functional proteins based solely on sequence data.

The key challenge of applying generative modeling to protein engineering is that the goal is usually not to generate diverse sequences that show similar properties to natural proteins, but to generate sequences that show improved properties. However, one application where the ability of generative modeling to design large numbers of diverse, functional sequences might be practically

useful is in library creation for the discovery of antibodies and other target-specific binders[115,116]. For example, Shin *et al*. recently applied this approach to design a smart library of diverse, stable and well-expressed nanobodies, using an autoregressive model trained on ~1.2 million naïve llama nanobody sequences[116]. A major advantage of autoregressive models in this context is that they use unaligned sequences rather than multiple sequence alignments; this is particularly useful in the case of antibody sequences, which have complementarity-determining regions of variable lengths and cannot be aligned accurately. A nanobody library containing ~185,000 highly diverse sequences generated by the model was constructed and experimentally characterized in a yeast display system. The designed library showed higher expression than a state-of-the-art combinatorial synthetic library based on position-specific amino acid frequencies, which was attributed to the ability of the design algorithm to account for higher-order sequence constraints. Another strategy for using generative models in protein engineering is to bias sequence generation towards sequences predicted to have some desirable property (e.g., stability or solubility)[112,115]. For example, Hawkins-Hooker *et al*. trained a conditional variational autoencoder on ~70,000 luciferase sequences labeled as low, medium, or high solubility based on a computational prediction[112]. Generation of predicted medium- and high-solubility variants of the *Pseudomonas aeruginosa* luciferase LuxA using this model yielded eight functional variants with reasonable solubility (>10% soluble expression) out of 23 variants tested, whereas the wild-type protein did not display soluble expression. However, this approach relies on the availability of an accurate sequence-based predictor for the property of interest. Although there are currently few experimentally validated examples where generative models have been used for protein engineering, it is a promising and rapidly developing area of research (recent theoretical developments are reviewed in ref. 108).

A different application of sequence data in machine learning-guided protein engineering is semi-supervised learning, where the performance of supervised learning with small training datasets can be improved using information learned from unlabeled protein sequence data (Fig. 3). To learn the relationship between protein sequence and fitness, supervised machine learning methods require the input protein sequences to be encoded as numeric vectors. This can be achieved using very basic representations of protein sequence (e.g., one-hot encoding, where the identity of the amino acid at each position is represented by a series of 19 zeroes and 1 one); however, the choice of representation is far from arbitrary, and more meaningful representations that summarize useful information about the protein sequences, for example, by representing closely related sequences in similar ways, can sometimes improve the performance of supervised learning[126]. As mentioned above, certain types of deep generative models trained on large sequence datasets are capable of learning meaningful, low-dimensional representations of protein sequence that capture information about amino acid properties and protein structure, phylogeny, and function[117,118]; these representations can therefore be useful for semi-supervised learning[121,126–129]. Using TEM-1 β-lactamase and *Aequorea victoria* green fluorescent protein (GFP) as model systems, Biswas *et al*. recently applied this approach to protein engineering by training an autoregressive model on the UniRef50 sequence database to learn the general features of protein sequences, then fine-tuning the model on sequences related to the target protein to learn the specific features of the target protein family[127]. The representations from these models were then used to perform supervised learning with a small training dataset (24 or 96 variants), and small libraries of improved variants were designed by performing *in silico* directed evolution using the resulting model. Up to 26% of the designed variants (depending on the protein and training set) showed improvements over the

wild-type function. Additionally, GFP variants were obtained that surpassed those obtained by ASR and consensus design (maximum fold increase in fluorescence compared with avGFP for each method: semi-supervised learning, 5.67; ASR, 2.51; consensus design, 2.47), and rivalled state-of-the-art variants previously obtained through a laborious, iterative engineering process (sfGFP, 6.52). The success of this method was attributed to the division of labor between the unsupervised model, which was able to broadly distinguish between functional and non-functional sequences, and the supervised model, which was then able to discriminate mediocre sequences from improved sequences.

An alternative to machine learning for sequence-based protein design is to use simpler statistical models such as those based on direct coupling analysis (DCA)[130–133]. This method uses a large multiple sequence alignment of a protein family to model the sequence probability distribution based on intrinsic amino acid propensities at each position and pairwise amino acid correlations at each pair of positions, thus explicitly accounting for covariation between residues. Russ *et al*. recently showed that this approach can be used to design enzymes with catalytic activity similar to natural proteins; when applied to a family of chorismate mutases, the designed enzymes showed a frequency of functional expression similar to natural enzymes (48% and 38% respectively) based on genetic complementation[131].

Concluding remarks

The methods described above allow reliable engineering or design, based on readily available data, of diverse proteins that show favorable properties such as high thermostability and expression, while minimizing the need for labor-intensive and expensive high-throughput

screening. Some of these methods, such as ASR and structure-based design, can also provide functionally diverse proteins, which may be useful of themselves in applications such as biocatalysis, or may be useful as robust and evolvable starting points for further engineering. Even as alternative methodologies such as *de novo* computational design and machine learning continue to improve, we expect that large-scale sequence data will remain a useful and complementary source of information in protein engineering. This trend is also seen in other fields of protein science; for example, an important factor in the breakthrough success of AlphaFold2[134] in the CASP14 structure prediction competition was the use of sophisticated deep learning methods to distill both structural and evolutionary data, including pairwise correlations encoded in multiple sequence alignments, together with physical and geometric restraints.

Sequence-based methods allow protein engineers to focus their exploration of sequence space on a limited region, loosely bounded by a set of homologous natural sequences provided in a multiple sequence alignment or training set (Fig. 4). This region of sequence space is only a small portion of the total sequence space associated with a particular fold or function[135], yet it encompasses an astronomical number of proteins that are functionally diverse and can greatly surpass natural proteins in terms of thermostability or other useful properties. Although the search strategy used in sequence-based protein engineering is conservative and may limit opportunities to engineer radically different functions[107], it has the major advantage of providing candidate sequences that have a high probability of being functional. This allows us to be adventurous in designing sequences that have a large number of mutations relative to any known natural protein sequence, and can thereby accelerate the pace of protein engineering. Directed evolution, for example, is limited in the number of mutations that can be introduced in a single round, because

random mutations tend on average to be destabilizing and have a negative impact on function, especially in combination[2–5]. However, there is a limit to what can be achieved with a single mutation; for example, an optimal single mutation identified by directed evolution usually yields a <10-fold improvement in catalytic activity, whereas >10 mutations are usually required to achieve a >10^3-fold improvement[97]. Likewise, only ~7% of stabilizing point mutations listed in the FireProtDB database were reported to increase T_m by >10 °C[136]. On the other hand, techniques like ASR, PROSS and FireProt routinely yield improvements in thermostability up to ~30 °C without sacrificing protein function, even while introducing tens or hundreds[49] of substitutions relative to any known natural protein sequence. Importantly, the ability of sequence-based design to introduce multiple mutations simultaneously also addresses the problem of epistasis, where a mutation may be beneficial only in combination with other mutations, and may therefore be discarded if the mutations are introduced one at a time[80,81,93,137].

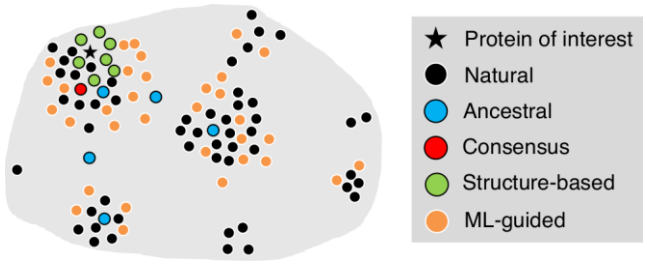


Figure 4. Schematic representation of the sequence space explored by different protein engineering and design methods. Using a large set of natural homologous sequences of a protein of interest, protein engineers can explore candidate sequences that have a high probability of being functional despite considerable sequence divergence from natural proteins. Sequences from consensus design, ASR, and structure-based approaches explore relatively close to clusters of natural sequences. Machine learning (ML)-guided approaches allow more systematic sampling of unexplored regions of sequence space further from natural sequences, although in practice,

sequences fairly similar to natural sequences are selected for characterization to eliminate false positives. This figure is based loosely on sequence space analyses from refs. 114 and 127.

Research at the interface of protein engineering and protein evolution has a long history of informing our understanding of the mechanisms and constraints of protein evolution while inspiring new protein engineering strategies[7,65,138]; for example, directed evolution has taught us about the importance of factors such as thermostability[139], promiscuity[140], and neutral drift[141] for protein evolvability. In the same way, sequence-based approaches to protein engineering and design are not only practically useful, but might extend our fundamental understanding of sequence-function relationships and the size, structure, and density of functional sequence space. For example, the design of functional proteins using coevolutionary methods such as DCA has shown that knowledge about amino acid conservation and pairwise correlations between amino acids is sufficient to specify protein structure and function[131,133]. These methods can also be used to estimate the maximum number of sequences in a protein family that possess the canonical fold or function[131,132,142]; for example, Russ *et al.* used their DCA model to estimate that ~10^{24} sequences in the AroQ family of chorismate mutases could be functional; this estimate is made more persuasive by the ability of the model to consistently generate functional proteins[131]. Thus, recent developments in generative modeling that enable systematic exploration of sequence space are likely to provide further insight into the fundamental question of how sequence encodes function.

We anticipate that sequence-based approaches will remain an important part of the protein engineering toolkit in the future and that their scope will be expanded by ongoing methodological

improvements. In machine learning, a key challenge is to develop and validate generative modeling approaches that can be used to design proteins with improved, rather than equivalent, properties compared with natural proteins. Various mathematical frameworks to achieve this by incorporating sequence-based predictors for properties such as stability or solubility are continuing to be developed and are awaiting experimental validation. When the property of interest cannot be predicted from sequence (e.g., catalytic specificity), semi-supervised learning approaches may be more useful, and it will be interesting to see how these approaches perform when applied more widely to non-model systems. There is also an ongoing need to develop tools that are accessible to non-experts in techniques such as structure-based design, phylogenetics, or machine learning; PROSS is a good example of an accessible tool that has been systematically validated across many laboratories. Finally, sequence databases are still growing at a rapid pace; improved methods for metagenomic analysis and strategic sequencing of organisms from underrepresented phylogenetic groups are continuing to provide useful, non-redundant protein sequences, which will extend the possibility of using sequence-based engineering methods to additional protein families.

AUTHOR INFORMATION

Corresponding Author

Paola Laurino (paola.laurino@oist.jp)

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

B.E.C. was supported by a JSPS Postdoctoral Fellowship for Overseas Researchers and a KAKENHI Grant-in-Aid for Scientific Research (20F20705) from the Japan Society for the Promotion of Science. Financial support from the Okinawa Institute of Science and Technology to P.L. is gratefully acknowledged.

ACKNOWLEDGMENT

We thank Sarel Fleishman and Devin Trudeau for critical reading of the manuscript.

ABBREVIATIONS

ASR, ancestral sequence reconstruction; DCA, direct coupling analysis; GFP, green fluorescent protein; ML, machine learning; MSA, multiple sequence alignment; WT, wild-type.

REFERENCES

(1) Ferguson, A. L., and Ranganathan, R. (2021) 100th anniversary of macromolecular science viewpoint: Data-driven protein design. *ACS Macro Lett. 10*, 327–340.

(2) Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov,

P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., Lukyanov, K. A., and Kondrashov, F. A. (2016) Local fitness landscape of the green fluorescent protein. *Nature 533*, 397–401.

(3) Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L., and Tawfik, D. S. (2007) The stability effects of protein mutations appear to be universally distributed. *J. Mol. Biol. 369*, 1318–1332.

(4) Guo, H. H., Choe, J., and Loeb, L. A. (2004) Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U.S.A. 101*, 9205–9210.

(5) Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., and Tawfik, D. S. (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature 444*, 929–932.

(6) Currin, A., Swainston, N., Day, P. J., and Kell, D. B. (2015) Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev. 44*, 1172–1239.

(7) Romero, P. A., and Arnold, F. H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol. 10*, 866–876.

(8) Packer, M. S., and Liu, D. R. (2015) Methods for the directed evolution of proteins. *Nat. Rev. Genet. 16*, 379–394.

(9) UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res. 49*, D480–D489.

(10) Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215*, 403–410.

(11) Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res. 39*, W29-W37.

(12) Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res. 49*, D412–D419.

(13) Blatt, L. M., Davis, J. M., Klein, S. B., and Taylor, M. W. (1996) The biologic activity and molecular characterization of a novel synthetic interferon-alpha species, consensus interferon. *J. Interferon Cytokine Res. 16*, 489–499.

(14) Steipe, B., Schiller, B., Plückthun, A., and Steinbacher, S. (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. *J. Mol. Biol. 240*, 188–192.

(15) Scrutton, N. S., Berry, A., and Perham, R. N. (1990) Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. *Nature 343*, 38–43.

(16) Onuffer, J. J., and Kirsch, J. F. (1995) Redesign of the substrate specificity of *Escherichia coli* aspartate aminotransferase to that of *Escherichia coli* tyrosine aminotransferase by homology modeling and site-directed mutagenesis. *Protein Sci. 4*, 1750–1757.

(17) Pan, X., and Kortemme, T. (2021) Recent advances in *de novo* protein design: Principles, methods, and applications. *J. Biol. Chem. 296*, 100558.

(18) Yang, K. K., Wu, Z., and Arnold, F. H. (2019) Machine-learning-guided directed evolution for protein engineering. *Nat. Methods 16*, 687–694.

(19) Mazurenko, S., Prokop, Z., and Damborský, J. (2019) Machine learning in enzyme engineering. *ACS Catal. 10*, 1210-1223.

(20) Frappier, V., and Keating, A. E. (2021) Data-driven computational protein design. *Curr. Opin. Struct. Biol. 69*, 63–69.

(21) Pinto, G. P., Corbella, M., Demkiv, A. O., and Kamerlin, S. C. L. (2021) Exploiting enzyme evolution for computational protein design. *Trends Biochem. Sci.* (in press)

(22) Faber, M. S., and Whitehead, T. A. (2019) Data-driven engineering of protein therapeutics. *Curr. Opin. Biotechnol. 60*, 104–110.

(23) Robinson, S. L., Piel, J., and Sunagawa, S. (2021) A roadmap for metagenomic enzyme discovery. *Nat. Prod. Rep. 38*, 1994–2023.

(24) Porebski, B. T., and Buckle, A. M. (2016) Consensus protein design. *Protein Eng. Des. Sel. 29*, 245–251.

(25) Bershtein, S., Goldin, K., and Tawfik, D. S. (2008) Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol. 379*, 1029–1044.

(26) Lehmann, M., Kostrewa, D., Wyss, M., Brugger, R., D'Arcy, A., Pasamontes, L., and van Loon, A. P. (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng. 13*, 49–57.

(27) Porebski, B. T., Keleher, S., Hollins, J. J., Nickson, A. A., Marijanovic, E. M., Borg, N. A., Costa, M. G. S., Pearce, M. A., Dai, W., Zhu, L., Irving, J. A., Hoke, D. E., Kass, I., Whisstock, J. C., Bottomley, S. P., Webb, G. I., McGowan, S., and Buckle, A. M. (2016) Smoothing a rugged protein folding landscape by sequence-based redesign. *Sci. Rep. 6*, 33958.

(28) Cirri, E., Brier, S., Assal, R., Canul-Tec, J. C., Chamot-Rooke, J., and Reyes, N. (2018) Consensus designs and thermal stability determinants of a human glutamate transporter. *eLife 7*, e40110.

(29) Takagi, H., Kozuka, K., Mimura, K., Nakano, S., and Ito, S. (2021) Design of a full-consensus glutamate decarboxylase and its application to GABA biosynthesis. *ChemBioChem*. (in press)

(30) Yao, H., Cai, H., and Li, D. (2020) Thermostabilization of membrane proteins by consensus mutation: a case study for a fungal Δ8-7 sterol isomerase. *J. Mol. Biol. 432*, 5162–5183.

(31) Sternke, M., Tripp, K. W., and Barrick, D. (2019) Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U.S.A. 116*, 11275–11284.

(32) Sternke, M., Tripp, K. W., and Barrick, D. (2020) The use of consensus sequence information to engineer stability and activity in proteins. In *Methods in Enzymology*, vol. 643; Elsevier Inc.; pp 149–179.

(33) Jones, B. J., Kan, C. N. E., Luo, C., and Kazlauskas, R. J. (2020) Consensus Finder web tool to predict stabilizing substitutions in proteins. In *Methods in Enzymology*, vol. 643; Elsevier Inc.; pp 129–148.

(34) Jones, B. J., Lim, H. Y., Huang, J., and Kazlauskas, R. J. (2017) Comparison of five protein engineering strategies for stabilizing an α/β-hydrolase. *Biochemistry 56*, 6521–6532.

(35) Risso, V. A., Gavira, J. A., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2014) Phenotypic comparisons of consensus variants versus laboratory resurrections of Precambrian proteins. *Proteins 82*, 887–896.

(36) Okafor, C. D., Pathak, M. C., Fagan, C. E., Bauer, N. C., Cole, M. F., Gaucher, E. A., and Ortlund, E. A. (2018) Structural and dynamics comparison of thermostability in ancient, modern, and consensus elongation factor *tus*. *Structure 26*, 118-129.

(37) Akanuma, S., Nakajima, Y., Yokobori, S., Kimura, M., Nemoto, N., Mase, T., Miyazono, K., Tanokura, M., and Yamagishi, A. (2013) Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci. U.S.A. 110*, 11067–11072.

(38) Nakano, S., Motoyama, T., Miyashita, Y., Ishizuka, Y., Matsuo, N., Tokiwa, H., Shinoda, S., Asano, Y., and Ito, S. (2018) Benchmark analysis of native and artificial NAD^+-dependent enzymes generated by a sequence-based design method with and without phylogenetic data. *Biochemistry 57*, 3722–3732.

(39) Spence, M. A., Kaczmarski, J. A., Saunders, J. W., and Jackson, C. J. (2021) Ancestral sequence reconstruction for protein engineers. *Curr. Opin. Struct. Biol. 69*, 131–141.

(40) Risso, V. A., Sanchez-Ruiz, J. M., and Ozkan, S. B. (2018) Biotechnological and protein-engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol. 51*, 106–115.

(41) Gaucher, E. A., Thomson, J. M., Burgan, M. F., and Benner, S. A. (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature 425*, 285–288.

(42) Gaucher, E. A., Govindarajan, S., and Ganesh, O. K. (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature 451*, 704–707.

(43) Risso, V. A., Gavira, J. A., Mejia-Carmona, D. F., Gaucher, E. A., and Sanchez-Ruiz, J. M. (2013) Hyperstability and substrate promiscuity in laboratory resurrections of Precambrian β-lactamases. *J. Am. Chem. Soc. 135*, 2899–2902.

(44) Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z.-M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T. J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J. M., Gaucher, E. A., and Fernandez, J. M. (2011) Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat. Struct. Mol. Biol. 18*, 592–596.

(45) Barruetabeña, N., Alonso-Lerma, B., Galera-Prat, A., Joudeh, N., Barandiaran, L., Aldazabal, L., Arbulu, M., Alcalde, M., De Sancho, D., Gavira, J. A., Carrion-Vazquez, M., and Perez-Jimenez, R. (2019) Resurrection of efficient Precambrian endoglucanases for lignocellulosic biomass hydrolysis. *Commun. Chem. 2*, 76.

(46) Gumulya, Y., Baek, J.-M., Wun, S.-J., Thomson, R. E. S., Harris, K. L., Hunter, D. J. B., Behrendorff, J. B. Y. H., Kulig, J., Zheng, S., Wu, X., Wu, B., Stok, J. E., De Voss, J. J., Schenk, G., Jurva, U., Andersson, S., Isin, E. M., Bodén, M., Guddat, L., and Gillam, E. M. J. (2018) Engineering highly functional thermostable proteins using ancestral sequence reconstruction. *Nat. Catal. 1*, 878–888.

(47) Hendrikse, N. M., Charpentier, G., Nordling, E., and Syrén, P.-O. (2018) Ancestral diterpene cyclases show increased thermostability and substrate acceptance. *FEBS J. 285*, 4660–4673.

(48) Babkova, P., Sebestova, E., Brezovsky, J., Chaloupkova, R., and Damborsky, J. (2017) Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. *ChemBioChem 18*, 1448–1456.

(49) Thomas, A., Cutlan, R., Finnigan, W., van der Giezen, M., and Harmer, N. (2019) Highly thermostable carboxylic acid reductases generated by ancestral sequence reconstruction. *Commun. Biol. 2*, 429.

(50) Nakano, S., Kozuka, K., Minamino, Y., Karasuda, H., Hasebe, F., and Ito, S. (2020) Ancestral L-amino acid oxidases for deracemization and stereoinversion of amino acids. *Commun. Chem. 3*, 181.

(51) Wilding, M., Peat, T. S., Kalyaanamoorthy, S., Newman, J., Scott, C., and Jermiin, L. S. (2017) Reverse engineering: transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chem. 19*, 5375–5380.

(52) Nakano, S., Minamino, Y., Hasebe, F., and Ito, S. (2019) Deracemization and stereoinversion to aromatic D-amino acid derivatives with ancestral L-amino acid oxidase. *ACS Catal. 9*, 10152–10158.

(53) Zakas, P. M., Brown, H. C., Knight, K., Meeks, S. L., Spencer, H. T., Gaucher, E. A., and Doering, C. B. (2017) Enhancing the pharmaceutical properties of protein drugs by ancestral sequence reconstruction. *Nat. Biotechnol. 35*, 35–37.

(54) Knight, K. A., Coyle, C. W., Radford, C. E., Parker, E. T., Fedanov, A., Shields, J. M., Szlam, F., Purchel, A., Chen, M., Denning, G., Sniecinski, R. M., Lollar, P., Spencer, H. T., Gaucher, E. A., and Doering, C. B. (2021) Identification of coagulation factor IX variants with enhanced activity through ancestral sequence reconstruction. *Blood Adv. 5*, 3333–3343.

(55) Hendrikse, N. M., Holmberg Larsson, A., Svensson Gelius, S., Kuprin, S., Nordling, E., and Syrén, P.-O. (2020) Exploring the therapeutic potential of modern and ancestral

phenylalanine/tyrosine ammonia-lyases as supplementary treatment of hereditary tyrosinemia. *Sci. Rep. 10*, 1315.

(56) Whitfield, J. H., Zhang, W. H., Herde, M. K., Clifton, B. E., Radziejewski, J., Janovjak, H., Henneberger, C., and Jackson, C. J. (2015) Construction of a robust and sensitive arginine biosensor through ancestral protein reconstruction. *Protein Sci. 24*, 1412–1422.

(57) Nicoll, C. R., Bailleul, G., Fiorentini, F., Mascotti, M. L., Fraaije, M. W., and Mattevi, A. (2020) Ancestral-sequence reconstruction unveils the structural basis of function in mammalian FMOs. *Nat. Struct. Mol. Biol. 27*, 14–24.

(58) Bailleul, G., Nicoll, C. R., Mascotti, M. L., Mattevi, A., and Fraaije, M. W. (2021) Ancestral reconstruction of mammalian FMO1 enables structural determination, revealing unique features that explain its catalytic properties. *J. Biol. Chem. 296*, 100221.

(59) Chaloupkova, R., Liskova, V., Toul, M., Markova, K., Sebestova, E., Hernychova, L., Marek, M., Pinto, G. P., Pluskal, D., Waterman, J., Prokop, Z., and Damborsky, J. (2019) Light-emitting dehalogenases: Reconstruction of multifunctional biocatalysts. *ACS Catal. 9*, 4810–4823.

(60) Trudeau, D. L., Kaltenbach, M., and Tawfik, D. S. (2016) On the potential origins of the high stability of reconstructed ancestral proteins. *Mol. Biol. Evol. 33*, 2633–2641.

(61) Wheeler, L. C., Lim, S. A., Marqusee, S., and Harms, M. J. (2016) The thermostability and specificity of ancient proteins. *Curr. Opin. Struct. Biol. 38*, 37–43.

(62) Williams, P. D., Pollock, D. D., Blackburne, B. P., and Goldstein, R. A. (2006) Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput. Biol. 2*, e69.

(63) Yunus, I. S., Palma, A., Trudeau, D. L., Tawfik, D. S., and Jones, P. R. (2020) Methanol-free biosynthesis of fatty acid methyl ester (FAME) in *Synechocystis sp*. PCC 6803. *Metab. Eng. 57*, 217–227.

(64) Devamani, T., Rauwerdink, A. M., Lunzer, M., Jones, B. J., Mooney, J. L., Tan, M. A. O., Zhang, Z.-J., Xu, J.-H., Dean, A. M., and Kazlauskas, R. J. (2016) Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J. Am. Chem. Soc. 138*, 1046–1056.

(65) Trudeau, D. L., and Tawfik, D. S. (2019) Protein engineers turned evolutionists - the quest for the optimal starting point. *Curr. Opin. Biotechnol. 60*, 46–52.

(66) Khersonsky, O., and Tawfik, D. S. (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem. 79*, 471–505.

(67) Wheeler, L. C., and Harms, M. J. (2021) Were ancestral proteins less specific? *Mol. Biol. Evol. 38*, 2227–2239.

(68) Nakano, S., Niwa, M., Asano, Y., and Ito, S. (2019) Following the evolutionary track of a highly specific L-arginine oxidase by reconstruction and biochemical analysis of ancestral and native enzymes. *Appl. Environ. Microbiol. 85*, e00459-19.

(69) Vialle, R. A., Tamuri, A. U., and Goldman, N. (2018) Alignment modulates ancestral sequence reconstruction accuracy. *Mol. Biol. Evol. 35*, 1783–1797.

(70) Aadland, K., and Kolaczkowski, B. (2020) Alignment-integrated reconstruction of ancestral sequences improves accuracy. *Genome Biol. Evol. 12*, 1549–1565.

(71) Clifton, B. E., Whitfield, J. H., Sanchez-Romero, I., Herde, M. K., Henneberger, C., Janovjak, H., and Jackson, C. J. (2017) Ancestral protein reconstruction and circular permutation

for improving the stability and dynamic range of FRET sensors. In *Methods in Molecular Biology*, vol. 1596; Elsevier Inc.; pp 71–87.

(72) Eick, G. N., Bridgham, J. T., Anderson, D. P., Harms, M. J., and Thornton, J. W. (2017) Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Mol. Biol. Evol. 34*, 247–261.

(73) Musil, M., Khan, R. T., Beier, A., Stourac, J., Konegger, H., Damborsky, J., and Bednar, D. (2021) FireProtASR: A web server for fully automated ancestral sequence reconstruction. *Brief. Bioinformatics 22*, 1–11.

(74) Huang, P.-S., Boyken, S. E., and Baker, D. (2016) The coming of age of *de novo* protein design. *Nature 537*, 320–327.

(75) Li, R., Wijma, H. J., Song, L., Cui, Y., Otzen, M., Tian, Y., Du, J., Li, T., Niu, D., Chen, Y., Feng, J., Han, J., Chen, H., Tao, Y., Janssen, D. B., and Wu, B. (2018) Computational redesign of enzymes for regio- and enantioselective hydroamination. *Nat. Chem. Biol. 14*, 664–670.

(76) Meng, Q., Capra, N., Palacio, C. M., Lanfranchi, E., Otzen, M., van Schie, L. Z., Rozeboom, H. J., Thunnissen, A.-M. W. H., Wijma, H. J., and Janssen, D. B. (2020) Robust ω-transaminases by computational stabilization of the subunit interface. *ACS Catal. 10*, 2915–2928.

(77) Aalbers, F. S., Fürst, M. J., Rovida, S., Trajkovic, M., Gómez Castellanos, J. R., Bartsch, S., Vogel, A., Mattevi, A., and Fraaije, M. W. (2020) Approaching boiling point stability of an alcohol dehydrogenase through computationally-guided enzyme engineering. *eLife 9*, e54639.

(78) Cui, Y., Wang, Y., Tian, W., Bu, Y., Li, T., Cui, X., Zhu, T., Li, R., and Wu, B. (2021) Development of a versatile and efficient C–N lyase platform for asymmetric hydroamination via computational enzyme redesign. *Nat. Catal. 4*, 364–373.

(79) Baker, D. (2019) What has de novo protein design taught us about protein folding and biophysics? *Protein Sci. 28*, 678–683.

(80) Weinstein, J., Khersonsky, O., and Fleishman, S. J. (2020) Practically useful protein-design methods combining phylogenetic and atomistic calculations. *Curr. Opin. Struct. Biol. 63*, 58–64.

(81) Khersonsky, O., Lipsh, R., Avizemer, Z., Ashani, Y., Goldsmith, M., Leader, H., Dym, O., Rogotner, S., Trudeau, D. L., Prilusky, J., Amengual-Rigo, P., Guallar, V., Tawfik, D. S., and Fleishman, S. J. (2018) Automated design of efficient and functionally diverse enzyme repertoires. *Mol. Cell 72*, 178-186.

(82) Netzer, R., Listov, D., Lipsh, R., Dym, O., Albeck, S., Knop, O., Kleanthous, C., and Fleishman, S. J. (2018) Ultrahigh specificity in a network of computationally designed protein-interaction pairs. *Nat. Commun. 9*, 5286.

(83) Bednar, D., Beerens, K., Sebestova, E., Bendl, J., Khare, S., Chaloupkova, R., Prokop, Z., Brezovsky, J., Baker, D., and Damborsky, J. (2015) FireProt: Energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol. 11*, e1004556.

(84) Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., and Damborsky, J. (2017) FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res. 45*, W393–W399.

(85) Pearce, R., Huang, X., Setiawan, D., and Zhang, Y. (2019) EvoDesign: Designing protein-protein binding interactions using evolutionary interface profiles in conjunction with an optimized physical energy function. *J. Mol. Biol. 431*, 2467–2476.

(86) Brender, J. R., Shultis, D., Khattak, N. A., and Zhang, Y. (2017) An evolution-based approach to *de novo* protein design. In *Methods in Molecular Biology.* vol. 1529; Elsevier Inc.; pp 243–264.

(87) Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R. L., Aharoni, A., Silman, I., Sussman, J. L., Tawfik, D. S., and Fleishman, S. J. (2016) Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell 63*, 337–346.

(88) Sumbalova, L., Stourac, J., Martinek, T., Bednar, D., and Damborsky, J. (2018) HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res. 46*, W356–W362.

(89) Broom, A., Jacobi, Z., Trainor, K., and Meiering, E. M. (2017) Computational tools help improve protein stability but with a solubility tradeoff. *J. Biol. Chem. 292*, 14349–14361.

(90) Kriegel, M., Wiederanders, H. J., Alkhashrom, S., Eichler, J., and Muller, Y. A. (2021) A PROSS-designed extensively mutated estrogen receptor α variant displays enhanced thermal stability while retaining native allosteric regulation and structure. *Sci. Rep. 11*, 10509.

(91) Lambert, A. R., Hallinan, J. P., Werther, R., Głów, D., and Stoddard, B. L. (2020) Optimization of protein thermostability and exploitation of recognition behavior to engineer altered protein-DNA recognition. *Structure 28*, 760-775.

(92) Barthel, S., Palluk, S., Hillson, N. J., Keasling, J. D., and Arlow, D. H. (2020) Enhancing terminal deoxynucleotidyl transferase activity on substrates with 3' terminal structures for enzymatic de novo DNA synthesis. *Genes (Basel) 11*, 102.

(93) Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., and Bradley, P. (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in Enzymology*, vol. 487; Elsevier Inc.; pp 545–574.

(94) Peleg, Y., Vincentelli, R., Collins, B. M., Chen, K.-E., Livingstone, E. K., Weeratunga, S., Leneva, N., Guo, Q., Remans, K., Perez, K., Bjerga, G. E. K., Larsen, Ø., Vaněk, O., Skořepa, O., Jacquemin, S., Poterszman, A., Kjær, S., Christodoulou, E., Albeck, S., Dym, O., and Fleishman, S. J. (2021) Community-wide experimental evaluation of the PROSS stability-design method. *J. Mol. Biol. 433*, 166964.

(95) Beerens, K., Mazurenko, S., Kunka, A., Marques, S. M., Hansen, N., Musil, M., Chaloupkova, R., Waterman, J., Brezovsky, J., Bednar, D., Prokop, Z., and Damborsky, J. (2018) Evolutionary analysis is a powerful complement to energy calculations for protein stabilization. *ACS Catal. 8*, 9420–9428.

(96) Bengel, L. L., Aberle, B., Egler-Kemmerer, A.-N., Kienzle, S., Hauer, B., and Hammer, S. C. (2021) Engineered enzymes enable selective *N*-alkylation of pyrazoles with simple haloalkanes. *Angew. Chem. Int. Ed. 60*, 5554–5560.

(97) Goldsmith, M., and Tawfik, D. S. (2017) Enzyme engineering: reaching the maximal catalytic efficiency peak. *Curr. Opin. Struct. Biol. 47*, 140–150.

(98) Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022) A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol. 23*, 40–55.

(99) Bryant, D. H., Bashir, A., Sinai, S., Jain, N. K., Ogden, P. J., Riley, P. F., Church, G. M., Colwell, L. J., and Kelsic, E. D. (2021) Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol. 39*, 691–696.

(100) Bedbrook, C. N., Yang, K. K., Robinson, J. E., Mackey, E. D., Gradinaru, V., and Arnold, F. H. (2019) Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nat. Methods 16*, 1176–1184.

(101) Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J., and Arnold, F. H. (2019) Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A. 116*, 8852–8858.

(102) Mason, D. M., Friedensohn, S., Weber, C. R., Jordi, C., Wagner, B., Meng, S. M., Ehling, R. A., Bonati, L., Dahinden, J., Gainza, P., Correia, B. E., and Reddy, S. T. (2021) Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng. 5*, 600–612.

(103) Fox, R. J., Davis, S. C., Mundorff, E. C., Newman, L. M., Gavrilovic, V., Ma, S. K., Chung, L. M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Whitman, J. C., Sheldon, R. A., and Huisman, G. W. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol. 25*, 338–344.

(104) Romero, P. A., Krause, A., and Arnold, F. H. (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci. U.S.A. 110*, E193–E201.

(105) Unger, E. K., Keller, J. P., Altermatt, M., Liang, R., Matsui, A., Dong, C., Hon, O. J., Yao, Z., Sun, J., Banala, S., Flanigan, M. E., Jaffe, D. A., Hartanto, S., Carlen, J., Mizuno, G. O., Borden, P. M., Shivange, A. V., Cameron, L. P., Sinning, S., Underhill, S. M., and Tian, L. (2020) Directed evolution of a selective and sensitive serotonin sensor via machine learning. *Cell 183*, 1986-2002.

(106) Wu, Z., Johnston, K. E., Arnold, F. H., and Yang, K. K. (2021) Protein sequence design with deep generative models. *Curr. Opin. Chem. Biol. 65*, 18–27.

(107) Wittmann, B. J., Johnston, K. E., Wu, Z., and Arnold, F. H. (2021) Advances in machine learning for directed evolution. *Curr. Opin. Struct. Biol. 69*, 11–18.

(108) Osadchy, M., and Kolodny, R. (2021) How deep learning tools can help protein engineers find good sequences. *J. Phys. Chem. B 125*, 6440–6450.

(109) Costello, Z., and Martin, H. G. (2019) How to hallucinate functional proteins. *arXiv:*1903.00458.

(110) Greener, J. G., Moffat, L., and Jones, D. T. (2018) Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep. 8*, 16189.

(111) Riesselman, A. J., Ingraham, J. B., and Marks, D. S. (2018) Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods 15*, 816–822.

(112) Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. (2021) Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol. 17*, e1008736.

(113) Gupta, A., and Zou, J. (2019) Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intell. 1*, 105–111.

(114) Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., Poviloniene, S., Laurynenas, A., Viknander, S., Abuajwa, W., Savolainen, O., Meskys, R., Engqvist, M. K. M., and Zelezniak, A. (2021) Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell. 3*, 324–333.

(115) Amimeur, T., Shaver, J. M., Ketchem, R. R., Taylor, J. A., Clark, R. H., Smith, J., Van Citters, D., Siska, C. C., Smidt, P., Sprague, M., Kerwin, B. A., and Pettit, D. (2020) Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. *bioRxiv*. DOI: 10.1101/2020.04.12.024844

(116) Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. (2021) Protein design and variant prediction using autoregressive generative models. *Nat. Commun. 12*, 2403.

(117) Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019) Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods 16*, 1315–1322.

(118) Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A. 118*, e2016239118.

(119) Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. (2020) ProGen: Language modeling for protein generation. *arXiv*:2004.03497.

(120) Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2021) ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv*:2007.06225.

(121) Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., and Song, Y. S. (2019) Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, vol. 32; Curran Associates, Inc.; pp 9689–9701.

(122) Hie, B., Zhong, E. D., Berger, B., and Bryson, B. (2021) Learning the language of viral evolution and escape. *Science 371*, 284–288.

(123) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017) Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30; Curran Associates, Inc.; pp 6000–6010.

(124) Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. (2021) Disease variant prediction with deep generative models of evolutionary data. *Nature 599*, 91–95.

(125) Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., and Naik, N. (2021) Deep neural language modeling enables functional protein generation across families. *bioRxiv*. DOI: 10.1101/2021.07.18.452833

(126) Yang, K. K., Wu, Z., Bedbrook, C. N., and Arnold, F. H. (2018) Learned protein embeddings for machine learning. *Bioinformatics 34*, 2642–2648.

(127) Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. (2021) Low-N protein engineering with data-efficient deep learning. *Nat. Methods 18*, 389–396.

(128) Luo, Y., Jiang, G., Yu, T., Liu, Y., Vo, L., Ding, H., Su, Y., Qian, W. W., Zhao, H., and Peng, J. (2021) ECNet is an evolutionary context-integrated deep learning framework for protein engineering. *Nat. Commun. 12*, 5743.

(129) Wittmann, B. J., Yue, Y., and Arnold, F. H. (2021) Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst. 12*, 1026–1045.

(130) Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R., and Weigt, M. (2018) Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys. 81*, 032601.

(131) Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. (2020) An evolution-based model for designing chorismate mutase enzymes. *Science 369*, 440–445.

(132) Trinquier, J., Uguzzoni, G., Pagnani, A., Zamponi, F., and Weigt, M. (2021) Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun. 12*, 5800.

(133) Tian, P., Louis, J. M., Baber, J. L., Aniana, A., and Best, R. B. (2018) Co-evolutionary fitness landscapes for sequence design. *Angew. Chem. Int. Ed. 57*, 5674–5678.

(134) Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., and

Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature 596*, 583–589.

(135) Povolotskaya, I. S., and Kondrashov, F. A. (2010) Sequence space and the ongoing expansion of the protein universe. *Nature 465*, 922–926.

(136) Stourac, J., Dubrava, J., Musil, M., Horackova, J., Damborsky, J., Mazurenko, S., and Bednar, D. (2021) FireProtDB: database of manually curated protein stability data. *Nucleic Acids Res. 49*, D319–D324.

(137) Markova, K., Chmelova, K., Marques, S. M., Carpentier, P., Bednar, D., Dambosrky, J., and Marek, M. (2020) Decoding the intricate network of molecular interactions of a hyperstable engineered biocatalyst. *Chem. Sci. 11*, 11162–11178.

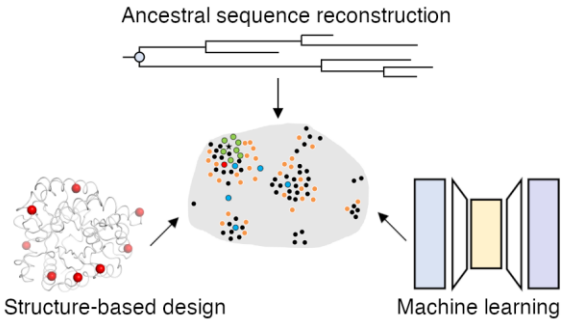
(138) Peisajovich, S. G., and Tawfik, D. S. (2007) Protein engineers turned evolutionists. *Nat. Methods 4*, 991–994.

(139) Bloom, J. D., Labthavikul, S. T., Otey, C. R., and Arnold, F. H. (2006) Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A. 103*, 5869–5874.

(140) Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., and Tawfik, D. S. (2005) The "evolvability" of promiscuous protein functions. *Nat. Genet. 37*, 73–76.

(141) Bloom, J. D., Romero, P. A., Lu, Z., and Arnold, F. H. (2007) Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct 2*, 17.

(142) Tian, P., and Best, R. B. (2017) How many protein sequences fold to a given structure? A coevolutionary analysis. *Biophys. J. 113*, 1719–1730.



For Table of Contents use only.