Pseudo-perplexity in one fell swoop for protein fitness estimation

Pranav Kantroo,[1,2,*] Günter P. Wagner,[3,4,5] and Benjamin B. Machta[2,6,†]
[1]*Computational Biology and Bioinformatics Program, Yale University, New Haven, Connecticut 06520, USA*
[2]*Quantitative Biology Institute, Yale University, New Haven, Connecticut 06520, USA*
[3]*Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA*
[4]*Department of Evolutionary Biology, University of Vienna, Djerassi Platz 1, A-1030 Vienna, Austria*
[5]*Hagler Institute for Advanced Studies, Texas A&M, College Station, Texas 77843, USA*
[6]*Department of Physics, Yale University, New Haven, Connecticut 06520, USA*

(Received 15 July 2024; accepted 25 July 2025; published 21 August 2025)

Protein language models trained on the masked language modeling objective learn to predict the identity of hidden amino acid residues within a sequence using the remaining observable sequence as context. They do so by embedding the residues into a high dimensional space that encapsulates the relevant contextual cues. These embedding vectors serve as an informative context-sensitive representation that not only aids with the defined training objective, but can also be used for other tasks by downstream models. We propose a scheme to use the embeddings of an unmasked sequence to estimate the corresponding masked probability vectors for all the positions in a single forward pass through the language model. This one fell swoop (OFS) approach allows us to efficiently estimate the pseudo-perplexity of the sequence, a measure of the model's uncertainty in its predictions, that can also serve as a fitness estimate. We find that ESM2 OFS pseudo-perplexity performs nearly as well as the true pseudo-perplexity at fitness estimation, and more notably it defines a new state of the art on the ProteinGym Indels benchmark. The strong performance of the fitness measure prompted us to investigate if it could be used to detect the elevated stability reported in reconstructed ancestral sequences. We find that this measure ranks ancestral reconstructions as more fit than extant sequences. Finally, we show that the computational efficiency of the technique allows for the use of Monte Carlo methods that can rapidly explore functional sequence space.

DOI: 10.1103/zhx7-hcmm

I. INTRODUCTION

The field of protein engineering is in the midst of a renaissance. Deep learning-based methods have enabled unprecedented strides towards addressing questions relating to the folding, design, and fitness prediction of protein sequences [1–4]. Large language models based on the transformer architecture lie at the heart of powering a host of such techniques [5]. Protein language models are typically composed of stacks of such transformer blocks, which are trained on data derived from large swaths of protein sequences to model the underlying data distribution [6,7]. The training data can take the form of raw sequences, alignments of related sequences, protein structures, and any combination of these modalities [6,8–10].

Just as sentences are different from random strings of words, proteins too are different from random strings of amino acid residues. As proteins are products of evolution, they adhere to what has been termed *Protein Grammar*—a conceptual anchor that connotes the rules governing the underlying generative process [11]. Our ability to design effective proteins hinges on being able to parse and characterize the dependencies that define this grammar. The sheer scale of the data, combined with the enormous number of tunable parameters, allows machine learning models to implicitly learn these operational principles [12]. For models that exclusively utilize raw sequences, two broad modeling paradigms have emerged towards this end: the masked language modeling objective and the next-token prediction objective [13].

The masked language modeling objective involves learning to predict the identity of hidden amino acid residues in a sequence using the remaining observable sequence as context [6,14]. The objective forces the model to assimilate both general and context-specific features within the input sequence to identify which residues fit best at the masked positions. A model trained on this objective is called an encoder model, since it transforms the residues in the input sequence into a set of high dimensional vectors that encapsulates its understanding of the given input. This set of high dimensional vectors is called the embedding, which is a task-agnostic, informative representation of the sequence that can be leveraged by downstream models to perform specific tasks, like protein folding [6,15,16].

The next-token prediction objective, on the other hand, requires learning to predict the identity of the next amino acid residue using the preceding sequence as context [17]. Models trained on this objective are called autoregressive models, since they can be used to generate novel sequences

*Contact author: pranav.kantroo@yale.edu
†Contact author: benjamin.machta@yale.edu

by iteratively sampling the next amino acid in a scheme where the output of the model at each step is fed back as the input for the next step [4,18–20]. The autoregressive sampling scheme illustrates how the nature of the training objective shapes the use-cases that these models can be applied to: autoregressive models are naturally well-suited for generative tasks in protein design.

Both these classes of models can be used to quantify the degree to which an input sequence adheres to the rules of protein grammar as learned by the models. This is accomplished through the use of various heuristics that transform the raw model output—a set of probability distributions on the amino acid alphabet—into a scalar quantity. Such scalar estimates that measure adherence to protein grammar have been empirically validated to be related to protein fitness: the more a sequence conforms to the rules of the learned grammar, the higher its fitness tends to be. As a consequence, language models have found extensive use in variant effect prediction, which entails assessing how mutations in a sequence affect its function [4,21–23].

Autoregressive models have excelled at this fitness estimation task due to their ability to score protein variants with all three kinds of mutations: substitutions, insertions, and deletions. Encoder models, however, have been limited in their scope to evaluating variants that differ only by substitutions [4,23]. The reason for this discrepancy can be attributed to the differences in the properties of the respective heuristics that are used to arrive at the fitness assessments. In this manuscript, we develop a simple extension that enables encoders to be used for fitness estimation of indels, and also allows for their wider use for generative tasks in protein design.

II. QUANTIFYING ADHERENCE TO PROTEIN GRAMMAR

In encoder models, the prediction for a specific masked position takes the form of a probability vector, where the dimensions of the vector denote the respective tokens in the model's vocabulary—amino acid residues in this case. The values in this vector represent the learned estimates for the probability that a given residue will stand in place of the mask token, conditioned on the context from the unmasked fraction of the sequence [Fig. 1(b)]. In this setting, the magnitude of the probability mass for a residue can be used as a proxy for the degree to which the residue fits in the sequence at that specific position. Consequently, residues with a higher probability mass should map to sequences that are more likely to be functional, and therefore have a higher fitness score. This intuition is formalized via the masked marginal heuristic to evaluate mutations [22].

Let M be the set of positions in a wild-type sequence S^{wt} where substitutions have been introduced to create a mutant sequence S^{mt}. The masked marginal (MM) heuristic dictates that the fitness F of the mutant sequence relative to the wild-type sequence can be scored as follows:

$$F_{\mathrm{MM}}(S^{\mathrm{mt}}|S^{\mathrm{wt}}) = \sum_{i\in M} \ln \frac{P\left(r_i = r_i^{\mathrm{mt}}\middle|S_{-M}\right)}{P\left(r_i = r_i^{\mathrm{wt}}\middle|S_{-M}\right)}, \tag{1}$$

where r_i is used to denote the ith amino acid residue in a sequence, and S_{-M} is the protein sequence masked at all the positions where the substitutions have been introduced. A positive score means that the mutant has a higher fitness than the wild-type sequence, and a negative score implies that the opposite is true. Note that this evaluation heuristic is based on a log-likelihood ratio, and is therefore by definition limited in its use to scoring sequences that differ by substitutions. This is because residues that have been inserted or deleted have no equivalent counterparts to compare against in this evaluation scheme.

Autoregressive models circumvent this issue by evaluating sequences using an altogether different heuristic. The heuristic relies on the idea that the fitness of a sequence is proportional to its likelihood as a whole, as measured by the language model: the more typical a sequence looks, the higher its fitness score should be [4,13]. The likelihood of a sequence S of length L can be expressed as a product of terms where each term only depends on the preceding context in the sequence:

$$P(S) = \prod_{i=1}^{L} P\left(r_i = r_i^S\middle|S_{<i}\right), \tag{2}$$

where $S_{<i}$ denotes the identity of the residues in the sequence up until the $(i-1)$th position.

Autoregressive models can efficiently compute each term in the factorized expression in a parallelized fashion, yielding the individual terms to calculate the log-likelihood of the sequence. Normalizing the log-likelihood by the sequence length enables fitness comparisons across sequences with different lengths, elucidating why the heuristic also works for indel mutations [24]. The fitness evaluation metric used in decoder models thus typically takes the form of the per-residue log-likelihood (LL) that can be directly used for fitness comparisons across closely related sequences in a reference-free manner: the higher the per-residue log-likelihood, the higher the fitness of the sequence,

$$F_{\mathrm{LL}}(S) = \frac{1}{L}\sum_{i=1}^{L} \ln\left(P\left(r_i = r_i^S\middle|S_{<i}\right)\right). \tag{3}$$

Autoregressive models, by virtue of their ability to score both substitutions and indels via the per-residue sequence log-likelihoods, have thus become the *de facto* standard for the fitness estimation task [13,23]. However, an equivalent analog in the form of a pseudo-log-likelihood, also exists for encoder models [25]. The per-residue pseudo-log-likelihood (PLL) of a sequence, measured via an encoder model, can be expressed as follows:

$$F_{\mathrm{PLL}}(S) = \frac{1}{L}\sum_{i=1}^{L} \ln\left(P\left(r_i = r_i^S\middle|S_{-i}\right)\right), \tag{4}$$

where S_{-i} denotes the sequence S masked at the position i.

In principle, the pseudo-log-likelihoods obtained from encoder models could be used as a fitness measure for indel mutations. We contend that the primary deterrent for their use in this task is that the expression is prohibitively slow to compute. This is because unlike autoregressive models, each of the L terms in the expression needs to be calculated one-at-a-time by introducing the mask token at each position in the sequence. The pseudo-log-likelihood computation

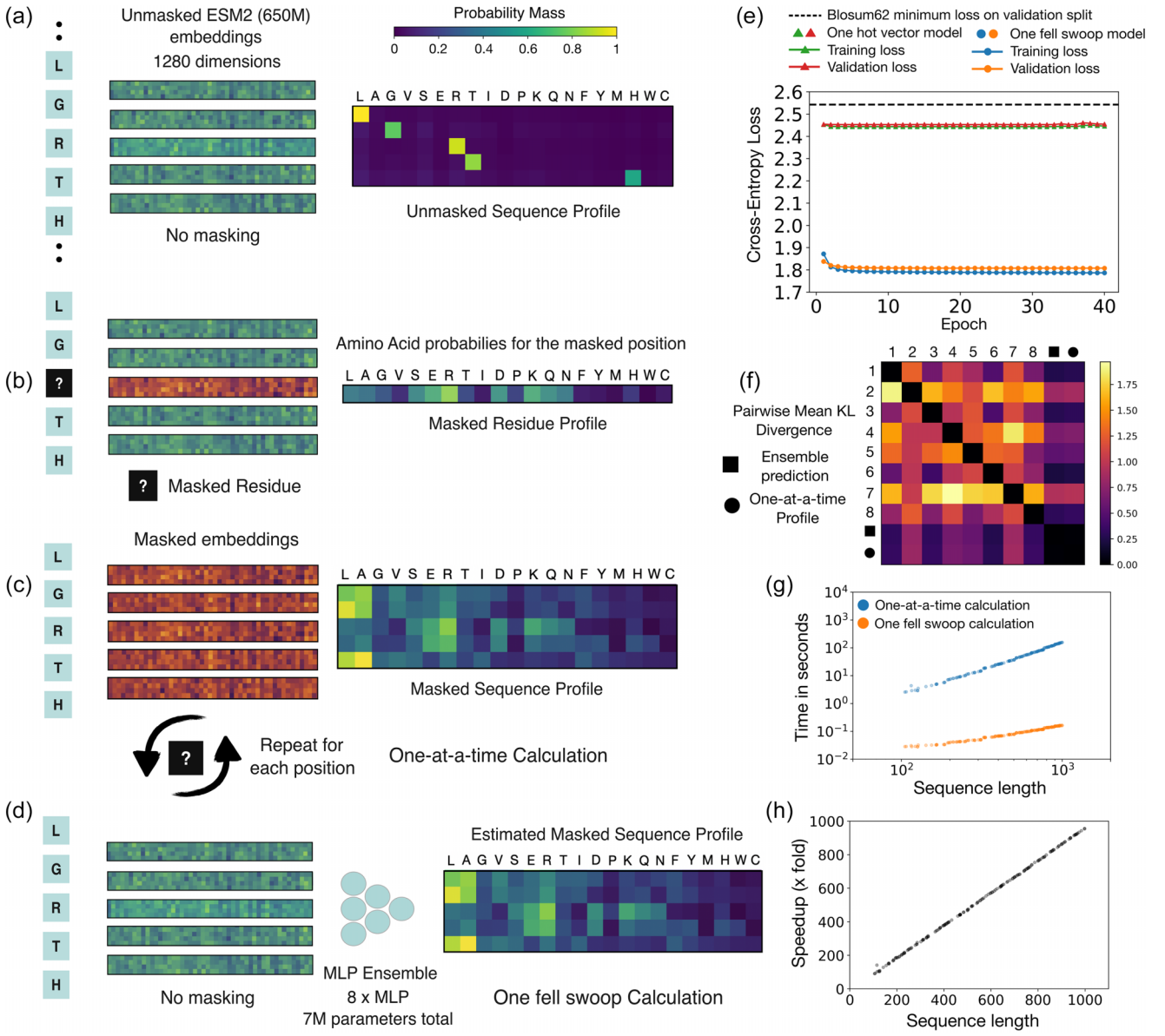

FIG. 1. The masked sequence profile can be calculated in one fell swoop. (a) The embeddings generated by the model take the form of a high dimensional vector—1280 dimensional in this case. The unmasked sequence profile of the input sequence is biased towards residues that are already present in the sequence, as seen by their large probability mass. (b) A representation of the masked residue profile for the sequence where the mask is placed at position 3. The embeddings of the unmasked residues are colored in green, while the embedding for the masked residue is orange. (c) A representation of the masked sequence profile calculation for the entire sequence where each residue is masked one-at-a-time to generate the corresponding masked residue profile. The sequence profile is obtained by collating the individual residue profiles into a matrix. (d) The one fell swoop setup, where the unmasked residue embeddings are used to estimate the corresponding masked residue profile for the position using an ensemble of eight multilayer perceptron models. The network is trained to minimize the cross-entropy loss between its prediction and the output of the language model. (e) A plot of the loss curves for the training and validation splits of the dataset to train the one fell swoop setup. As a baseline, we use a model that uses one hot vector representations of the amino acid sequence to estimate its masked sequence profile. We also plot the minimum loss (validation split) achieved by using the Blosum62 profile as logits to estimate the masked sequence profile. The temperature of the softmax operation that converts the logits to probabilities is chosen to minimize the loss for the validation split. (f) We use pairwise mean KL divergence between the substitution profile predictions of the individual MLP models in the ensemble to visualize their heterogeneity. These predictions are made for sequences in the validation split. We also include predictions from the ensembled setup and the one-at-a-time profile in this comparison as a reference. (g) Plot of the time taken for the one-at-a-time calculation and the one fell swoop calculation vs sequence length using an A5000 GPU averaged over 10 runs. (h) Plot of the speedup as a function of sequence length when using the one fell swoop calculation.

thus consumes L forward passes through the model for a sequence of length L [Fig. 1(c)] [25]. In comparison, the masked-marginal strategy and the autoregressive parallelized log-likelihood computation both take just one forward pass regardless of sequence length.

III. PSEUDO-PERPLEXITY IN ONE FELL SWOOP

Pseudo-perplexity is a quantity that is derived from the per-residue pseudo-log-likelihood, and it measures the uncertainty in an encoding model's predictions for some given input. A low pseudo-perplexity indicates that the input sequence is in line with the assessment of the model based on what it learned from the data on which it was trained. It therefore represents how natural or typical the sequence seems to the model: the lower the pseudo-perplexity, the more natural the sequence looks. Pseudo-perplexity (PP) Z is inversely related to fitness, and its negative value can serve as a fitness estimate. It can be expressed in terms of the pseudo-log-likelihood of a sequence via the following expression [25]:

$$Z(S) = e^{-F_{\mathrm{PLL}}(S)}. \tag{5}$$

We get around the length dependency of the pseudo-log-likelihood calculation by proposing a one fell swoop (OFS) approach that can make do with just a single forward pass through the encoder model. The approach relies on the idea that the embedding of an unmasked residue contains enough information to be able to accurately infer the residue's masked profile. If this is true, then it should be possible to obtain the masked residue profile of the ith amino acid, $P(r_i = r_i^S|S_{-i})$, by projecting it from its unmasked embedding vector E_i using some learned function f:

$$f(E_i) \sim P\left(r_i = r_i^S \middle| S_{-i}\right). \tag{6}$$

Such a projection function from the embedding space to the space of probability vectors should enable single pass parallelized estimation of the pseudo-log-likelihood of the sequence. We thus reframe the pseudo-log-likelihood calculation as a regression problem to estimate the entire sequence profile using its unmasked residue embeddings. The function takes the form of a neural network model, and the training data to learn the mapping are generated by performing the one-at-a-time calculation for a large corpus of sequences.

We train an ensemble of eight multilayer perceptron models to generate the ESM2 (650M) sequence profile of an input sequence using its ESM2 (650M) unmasked residue embeddings [6] obtained from the last layer. The sequences used in this step are the representative sequences for protein clusters obtained by clustering a protein sequence database via mmseqs2 [26]. The model is trained to minimize the cross-entropy loss between its prediction and the profile matrix obtained via the one-at-a-time calculation [Figs. 1(c) and 1(d)]. The small value of the KL divergence between the predictions of the ensembled model and the sequence profile calculated via the one-at-a-time procedure [Fig. 1(f)] indicates that the one fell swoop method can be effectively used to approximate the sequence profile. This estimate is also reached in a fraction of the time it takes to perform the one-at-a-time calculation [Figs. 1(g) and 1(h)].

The method as outlined above can be used for any encoder model, however its effectiveness is contingent on the quality of the embeddings being used, and the relative difficulty of the regression task. The reason for the boost in efficiency can be pinned to the fact that the unmasked embeddings can be calculated in a single forward pass through the encoder model, and the subsequent profile calculation is performed in parallel by the much tinier one fell swoop setup—7M parameters in total. Calculating the pseudo-log-likelihoods of the sequence can thus be turned into a single pass affair, while also generating the entire single residue substitution-based mutational landscape of the sequence in the process. Note that the one fell swoop procedure produces an estimate of the ESM2 (650M) sequence profile, and not its exact value. Pseudo-perplexity calculated using this estimated sequence profile is denoted by OFS pseudo-perplexity to draw the distinction.

IV. EVALUATING OFS PSEUDO-PERPLEXITY AS A FITNESS ESTIMATOR

The reliability of computational fitness estimates can be assessed by comparing them against experimentally measured fitness scores from deep mutational scanning (DMS) experiments [27]. The degree of concordance between the language model fitness estimates and the experimentally determined ground truth can be used as a performance metric to assess how well these models do on the zero-shot fitness estimation task.

The ProteinGym benchmark is a compilation of data obtained from a large set of DMS experiments that spans a wide variety of selection assays: Activity, Binding, Expression, Organismal Fitness, and Stability. The extensive dataset can thus be used to benchmark the performance of these models and systematically document methodological improvements [23]. We tested the efficacy of ESM2 OFS pseudo-perplexity at zero-shot fitness prediction using experimentally determined fitness measurements drawn from DMS experiments and compared its performance with other baselines using data sourced from the ProteinGym repository. The consistency between the experimentally measured fitness scores and the language model assessment is measured using the Spearman rank correlation. The metric measures the correlation between the rank orderings of the fitness scores from the two sets.

We assessed the reliability of the OFS pseudo-perplexity fitness estimate by comparing its performance against pseudo-perplexity obtained via the one-at-a-time calculation for 61 DMS assays. The sum of the lengths of the variants in the 61 DMS assays is less than the chosen threshold of 40 000—enabling us to perform the full one-at-a-time calculation for the subset. We observe that the performance of the two measures is largely consistent, with true pseudo-perplexity having a slight edge over OFS pseudo-perplexity [Fig. 2(a)]. We also found that OFS pseudo-perplexity as an evaluation strategy incurs a modest decline in performance on the substitutions benchmark, as compared to the masked marginal heuristic (Table I). This observation is consistent with previously reported results that assessed the performance of pseudo-perplexity [22], and it shows that the masked marginal heuristic is the preferable evaluation strategy for ESM2 when exclusively dealing with

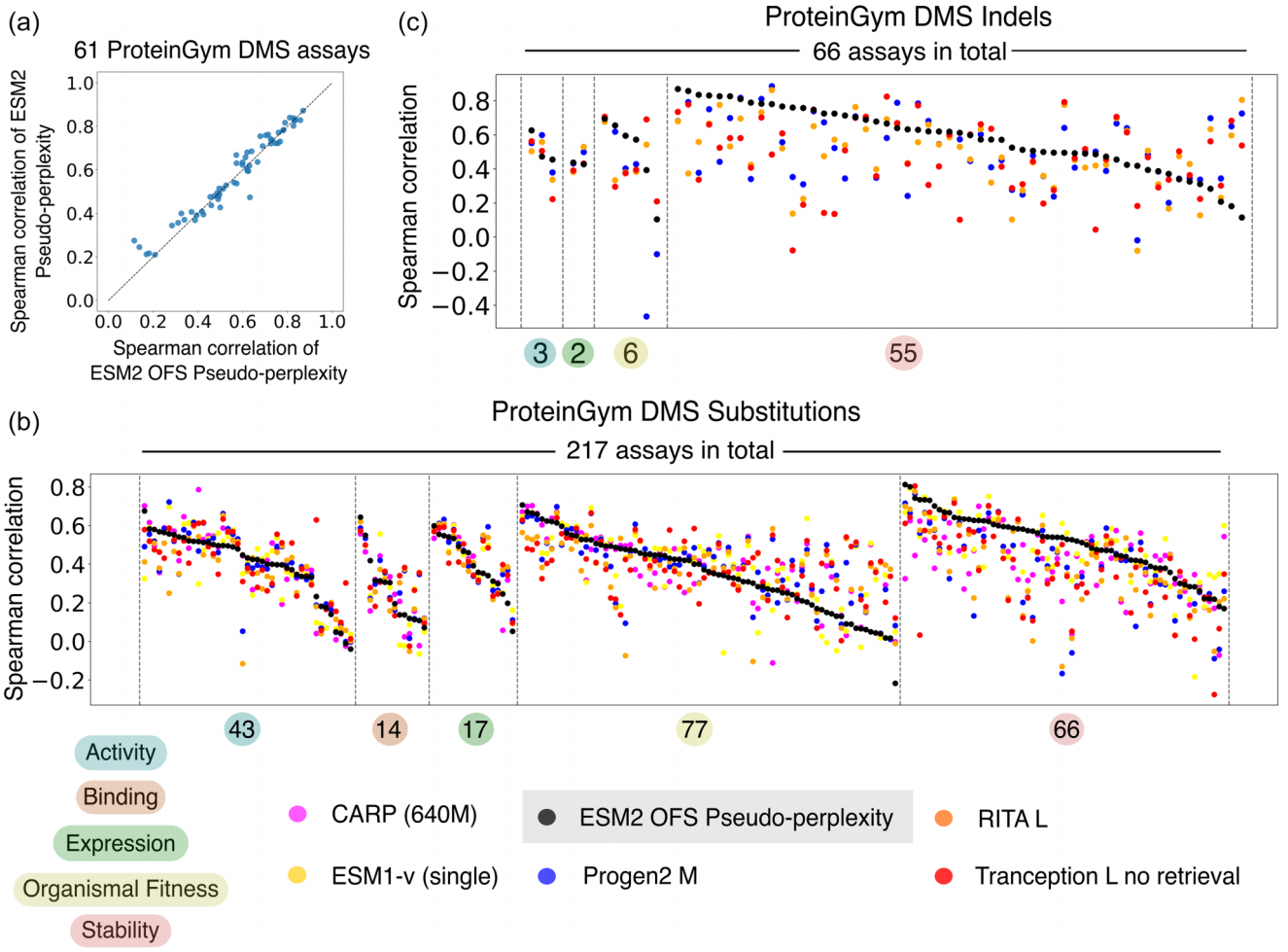

FIG. 2. OFS pseudo-perplexity is competitive with state-of-the-art methods for evaluating protein fitness. (a) The performance of the one fell swoop calculation at the fitness estimation task is similar to the performance of the one-at-a-time calculation. (b) The benchmarking data are arranged by function types, and the individual DMS experiments within each category are arranged in descending order of the performance of ESM2 OFS pseudo-perplexity as measured by the Spearman correlation. ESM2 OFS pseudo-perplexity is competitive with other state-of-the-art protein language models in its size class that exclusively utilize sequence information. The performance data for the other models were obtained from the ProteinGym repository. (c) ESM2 OFS pseudo-perplexity defines a new state of the art on the ProteinGym Indels benchmark.

substitutions. Comparison with other baselines indicates that ESM2 OFS pseudo-perplexity as a stand-alone fitness estimator is competitive with state-of-the-art protein language models in its size class that rely exclusively on sequence information [Fig. 2(b) and Table I]. The utility of the approach at fitness estimation is more apparent on the indels benchmark, where encoding models have not been tested thus far. Not only does OFS allow encoding models like ESM2 to be used as fitness estimators for insertions and deletions, we find that ESM2 OFS pseudo-perplexity defines a new state of the art on the indels benchmark [Fig. 2(c) and Table II].

We note that we have chosen to compare ESM2 only against similarly sized models that have been trained exclusively on raw protein sequences. Hybrid models can utilize additional data modalities like multiple sequence alignments or structure, and they are known to exhibit significant overall improvements in their benchmark performance compared to models that are trained exclusively on sequences [4,8,28–30]. The trend makes intuitive sense because sequence alignments and structure are richer modalities that explicitly contain information about the evolutionary context and the three-dimensional configuration that the protein acquires in space. This additional information in the form of evolutionarily constrained and flexible positions, the empirical distribution of viable residues at a specific position, or the physical constraints imposed due to structure can allow the model to identify broader trends and refine its assessment.

Notably, it has also been shown that models trained on specific data modalities can excel at capturing particular aspects of protein fitness. Specifically, inverse-folding models, which predict sequence likelihoods by using the structure of the sequence as the conditioning variable, along with hybrid models that incorporate structural information, have been found to be especially adept at fitness estimation for stability-based assays within ProteinGym [28–30]. Data drawn from specific selection assays can thus be used to conceptualize the model assessment in terms of the aspect of protein fitness that it best correlates with. In this context, we note that ESM2 OFS

TABLE I. ProteinGym Substitutions: Spearman correlation. The data for the baselines CARP, ESM-1v, Progen2 M, RITA L, and Tranception L no retrieval are sourced from the ProteinGym repository.

	Function type					
Language model	Activity (43)	Binding (14)	Expression (17)	Organismal fitness (77)	Stability (66)	Aggregate mean
ESM2: MM	0.430	0.332	0.440	0.363	0.518	0.430
ESM2: OFS PP	0.393	0.279	0.397	0.331	0.507	0.403
CARP (640M)	0.395	0.274	0.419	0.364	0.414	0.385
ESM-1v (single)	0.390	0.268	0.431	0.362	0.476	0.405
Progen2 M	0.393	0.296	0.436	0.381	0.396	0.388
RITA L	0.359	0.291	0.422	0.374	0.383	0.375
Tranception L no retrieval	0.401	0.289	0.415	0.389	0.381	0.386

pseudo-perplexity is most strongly correlated with data drawn from stability-based assays, as is evident from both the substitutions and the indels benchmark (Tables I and II). This observation led us to our next question: Can the fitness measure be used to capture the widely reported elevated stability of reconstructed ancestral sequences [31]?

V. ANCESTRAL STATE RECONSTRUCTION: OLD IS GOLD?

Sequence alignments of related extant sequences can be used to infer their phylogenetic relationship, and to study the functional evolution of the family over time through the topology of the inferred tree [32]. The topology along with the alignment data can in turn be used to estimate the states of the ancestral nodes in the tree through various statistical frameworks, in a process called ancestral state reconstruction [33]. The technique has made it possible to resurrect proteins that may have existed eons ago, and thereby directly investigate the functional properties of these inferred ancestral sequences [34–39]. The research program has led to the rather perplexing finding that reconstructed ancestral sequences tend to have superior functional properties compared to their extant counterparts: improved activity, enhanced stability, and/or higher promiscuity [40–43]. This is true, so much so that ancestral state reconstruction has emerged as a viable strategy for protein design [44,45].

The phenomenon as a whole is not well understood. One line of reasoning to explain the trend is that the technique suffers from an inherent bias in the form of a consensus effect, which systematically leads to sequences with a higher stability. The idea is that if a given residue is predominantly found in a specific position in a sequence alignment, then it is likely to be assigned as the ancestral state for the reconstructed ancestors. This reasoning is well-founded for positions that are constrained by function; however, it may also lead to spurious inference that is biased towards the consensus for positions where this is not true [31]. Replacing rare residues by the consensus residues drawn from the sequence alignment has also been shown to generate sequences with higher stability [46–49]. In this view, the unusual properties of ancestral sequences can be traced back to this bias towards the consensus.

An alternative line of reasoning for the phenomenon asserts that the inferred ancestral sequences appear to possess superior functional properties because the true ancestral sequences did in fact possess those properties. The functional properties of a protein are a product of the environment in which it operates. So, only proteins with a superior stability could have been functional in a pre-Cambrian hot bath, or a higher oxidative stress environment, or the higher radiation levels of a more tumultuous Earth of the past [31,41]. Furthermore, there have been proposals where subfunctionalization over evolutionary timescales can mediate a generalist-to-specialist transition, wherein promiscuous ancestral proteins progressively become more specialized in function [42,43,50,51]. This transition can also be used to argue for systematic differences between the two sets; however, a consensus for the existence of such a trend has also not been established [51].

Disentangling the confounding variables in this setup has not been straightforward. Regardless of the underlying reason, we set out to investigate if we can detect the disparity in function between extant and reconstructed ancestral sequences. Addressing this problem entails building a framework to distinguish between the statistical behavior of populations of sequences. We use ESM2 OFS pseudo-perplexity as a scalar

TABLE II. ProteinGym Indels: Spearman correlation. The data for the baselines CARP, ESM-1v, Progen2 M, RITA L, and Tranception L no retrieval are sourced from the ProteinGym repository.

	Function type				
Language model	Activity (3)	Expression (2)	Organismal fitness (6)	Stability (55)	Aggregate mean
ESM2: OFS PP	0.518	0.433	0.553	0.582	0.574
Progen2 M	0.510	0.466	0.374	0.517	0.503
RITA L	0.466	0.456	0.419	0.495	0.486
Tranception L no retrieval	0.431	0.412	0.445	0.469	0.464

proxy for function to assign the fitness scores for individual sequences. We then use Cliff's delta, a nonparametric estimator of effect size, to compare the fitness scores of the respective populations [54]. Cliff's delta measures the tendency of elements in A to be greater than those in B without making any assumptions about the distributions from which the values are drawn. For $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, where the elements correspond to the fitness scores for the respective sequences in the two sets in this setting, Cliff's delta (δ) is given by

$$\delta(A, B) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{sgn}(a_i - b_j)}{mn}, \tag{7}$$

where *sgn* denotes the signum function, which outputs $1, 0, -1$ based on the input's sign. Cliff's delta between the fitness scores of the two sets of sequences can thus be used as a measure for the disparity in function between the populations.

We calibrated this metaphorical instrument using data from ProteinGym DMS assays and found that the consistency in the sign of the inferred and true value of Cliff's delta strongly depends on the magnitude of the inferred value: the larger the magnitude, the more likely it is that the sign is inferred correctly [Fig. 3(a)]. The true value in this context is defined to be the negative Cliff's delta between the experimentally measured fitness scores of the sequences in the two sets. The negative sign is introduced in the expression to match its sign with the inferred disparity that uses the ESM2 OFS pseudo-perplexity of the sequences in the calculation. This observation prompted us to use a symmetric decision boundary of 0.33 around 0 to reliably ascertain the direction of the discrepancy in function. In this scheme, values that lie within the boundary are discarded as error-prone, and therefore uncertain, while values that lie outside of it can be used as confident estimates for the sign of the disparity in function [Figs. 3(c) and 3(d)] (see the Decision Boundary section in the Appendix for more details).

With a framework in place, we can return to addressing the original question: Can we capture the widely reported elevated stability of reconstructed ancestral sequences? Towards this end, we procured sequence alignments for 257 protein families from the Interpro database [55]. The pfam families [56] used in this analysis were not chosen randomly; however, we were cognizant so as not to introduce a systematic bias that would influence the outcome of the experiment (see the Ancestral State Reconstruction section in the Appendix for details). We used IQTREE2 [57,58] to infer the phylogeny of the families, and FastML [59] is used for ancestral state reconstruction. A positive value for Cliff's delta in this setting indicates that the extant sequences tend to have a higher ESM2 OFS pseudo-perplexity compared to the reconstructed sequences, while a negative value indicates the opposite trend. We found that Cliff's delta has a positive value greater than the chosen threshold of 0.33 for 205 out of 257 families (79.76%), and a negative value less than –0.33 for 5 out of 257 families (1.94%) as shown in Fig. 3. We note that some of the families with the negative value of Cliff's delta are polyphyletic in origin, which may explain the higher pseudo-perplexity of the reconstructed ancestors. The remaining 47 families lie in the uncertain range where estimates for the sign of Cliff's delta are not reliable.

The results imply that reconstructed ancestral sequences are predicted to have a higher stability for the majority of protein families under consideration as per the language model assessment. This is because a lower ESM2 OFS pseudo-perplexity is predictive of a higher stability as seen by the performance of the fitness measure on DMS assays.

It is tempting to speculate whether a lower pseudo-perplexity in the reconstructed sequences implies that a consensus effect is at play in the reconstruction inference process, since pseudo-perplexity does relate to typicality. Establishing and quantifying such a bias in the reconstruction procedure requires a more careful analysis, which we will pursue in an upcoming venture. For now, we only submit that the model's assessment can be used to detect the known disparity in function between the two sets of sequences. Ascribing a reason for precisely why this is the case is not straightforward, but the fact that the language model is in tune with this trend is intriguing. These observations lend additional support to the idea that large language models are more than just stochastic parrots [60].

VI. EXPLORING PROTEIN MORPHOSPACE

Protein design constitutes a broad range of design goals, such as enhancing the functional properties of existing sequences, modifying known sequences to acquire new functions, or even generating altogether novel sequences with as-yet unseen folds [61]. The common thread in the endeavor is to wade through protein morphospace and sample sequences that best satisfy some user-defined constraints. Machine learning models have been used to navigate through this vast space by pruning it down to the much smaller set of presumably viable states—states that have a high likelihood as per the model's assessment. Several such deep learning powered workflows have emerged that have been successfully deployed to generate functional protein sequences [7,11,62,63].

The design goals in such generative tasks can be broadly framed as *de novo* sequence generation, family-specific sequence generation, or structure-conditioned generation. Autoregressive language models are particularly well-suited for the *de novo* sequence generation task. These models utilize an autoregressive sampling scheme where the sequence manifests one residue at a time, and can be used to sample sequences that are far from natural proteins in sequence identity. Family-specific sequence generation can be performed by fine-tuning or training an autoregressive model on sequences drawn from the family of interest [7,64]. Alternatively, hybrid language models offer a fine-tuning-free way to generate family-specific sequences by using the sequences from the family to condition the generation process [65,66]. Likewise, inverse folding models have enabled structure-oriented sequence generation as they can model the distribution of protein sequences that acquire some given input structure [2,9].

In contrast to the above generative schemes, an iterative refinement workflow involves starting with a seed sequence that is progressively morphed to better fit the provided design constraints. Such workflows typically employ a Markov chain Monte Carlo framework for the refinement procedure. The design objective is framed as an energy minimization prob-

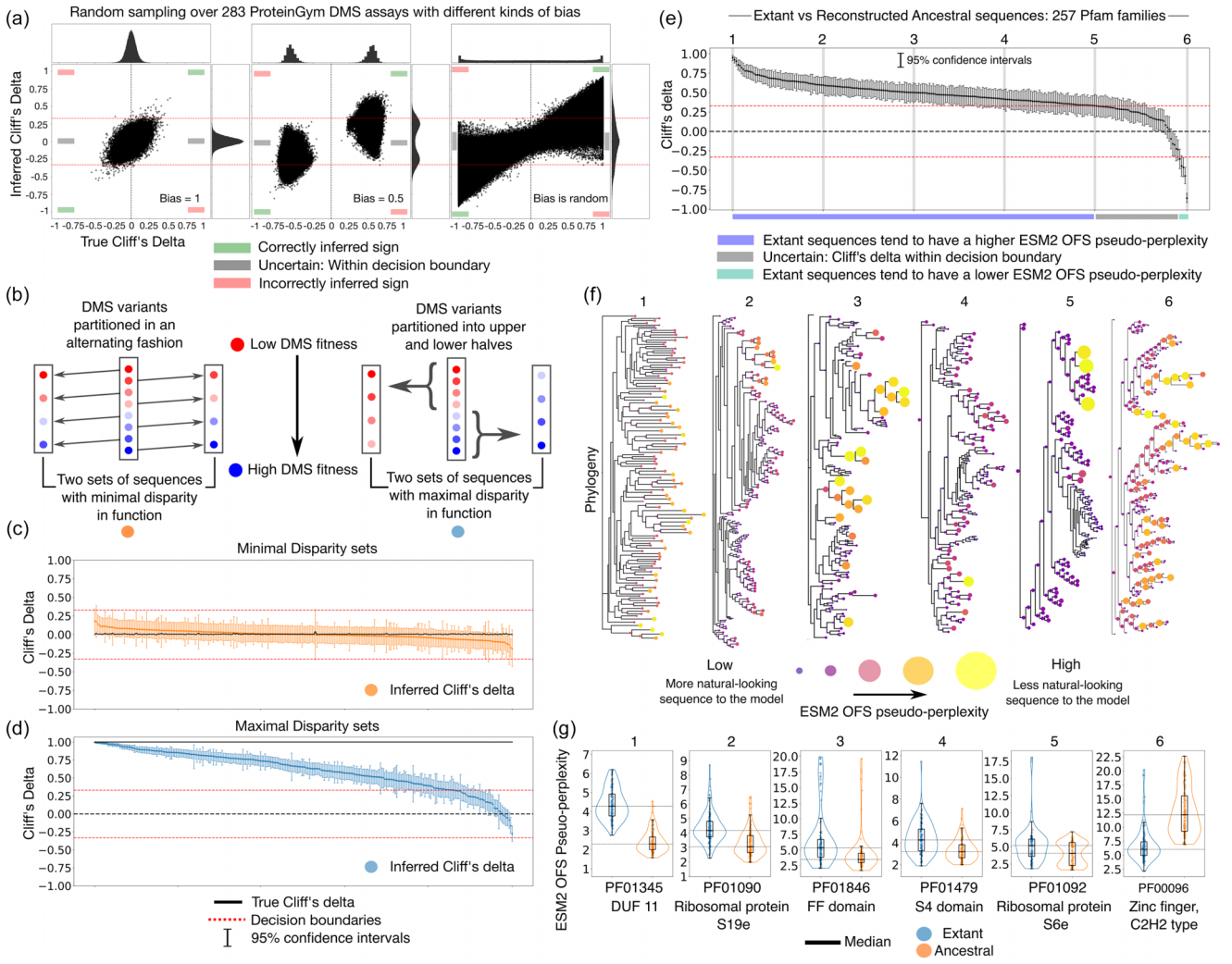

FIG. 3. Ancestral reconstructions are assigned higher fitness than extant proteins by OFS pseudo-perplexity. (a) Variants in 283 ProteinGym DMS assays are partitioned into two sets of sequences via random sampling, where each set has 200 or fewer sequences. Bias is introduced in the sampling process to generate populations with a larger disparity in function. A bias parameter of 1 corresponds to unbiased sampling, while a lower value induces a bias in the sampling process, generating populations with a larger disparity in function. We calculate the true and inferred value of Cliff's delta for the populations, and we observe that inferred values that are larger than 0.33 in magnitude (red dotted lines) are largely inferred correctly. (b) The minimal disparity sets are created by rank ordering the variants in ProteinGym DMS assays by their experimentally measured fitness scores and then partitioning them by selecting sequences alternately, while the maximal disparity sets are created by partitioning the variants into the upper and lower halves of the fitness spectrum. For assays where the total number of variants is greater than 400, only the top and bottom 200 sequences arranged by the fitness scores are used to generate the sets, while all the sequences are used for assays with fewer than 400 variants. (c) The Cliff's delta curve for the minimal disparity sets lies within the decision boundary indicating that the measured disparity between the sets is too small for its direction to be reliably inferred. (d) For the maximal disparity sets, we see that the procedure correctly ascertains the direction of the discrepancy in function for a significant proportion of pairs 243 out of 283 (85.86%), however the remaining 40 pairs do end up in the gray zone of uncertainty. (e) A plot of Cliff's delta for the ESM2 OFS pseudo-perplexity between extant and reconstructed ancestral sequences for 257 protein families. The families are arranged in descending order of the Cliff's delta between the two sets. The gray vertical bars are used to demarcate the families shown in detail in (f) and (g). The dotted line in red denotes the decision boundary. (f) Phylogeny of the respective families that correspond to the gray bars in (e). The interior nodes (reconstructed ancestral sequences) in the tree tend to have lower ESM2 OFS pseudo-perplexity values for the five trees that have a positive Cliff's delta, while the leaf nodes (extant sequences) tend to have a lower ESM2 OFS pseudo-perplexity in the last tree where the Cliff's delta value is negative. The size and color [52,53] of the pseudo-perplexity markers (solid circles) are only meant to be compared within a tree, and not across different trees. (g) The distribution of the fitness scores in the two sets of sequences visualized in a box plot for the six families marked by the gray bars in (e).

lem, where the energy function encodes the constraints that a successful design needs to satisfy: sequences that best satisfy the constraints have the lowest energy. The protocol entails sampling mutation proposals from some designated sampling distribution that can be accepted or rejected based on the energy of the resultant sequence. The optimization involves making such proposals in an iterative fashion to improve the design over time, yielding low-energy sequences that satisfy

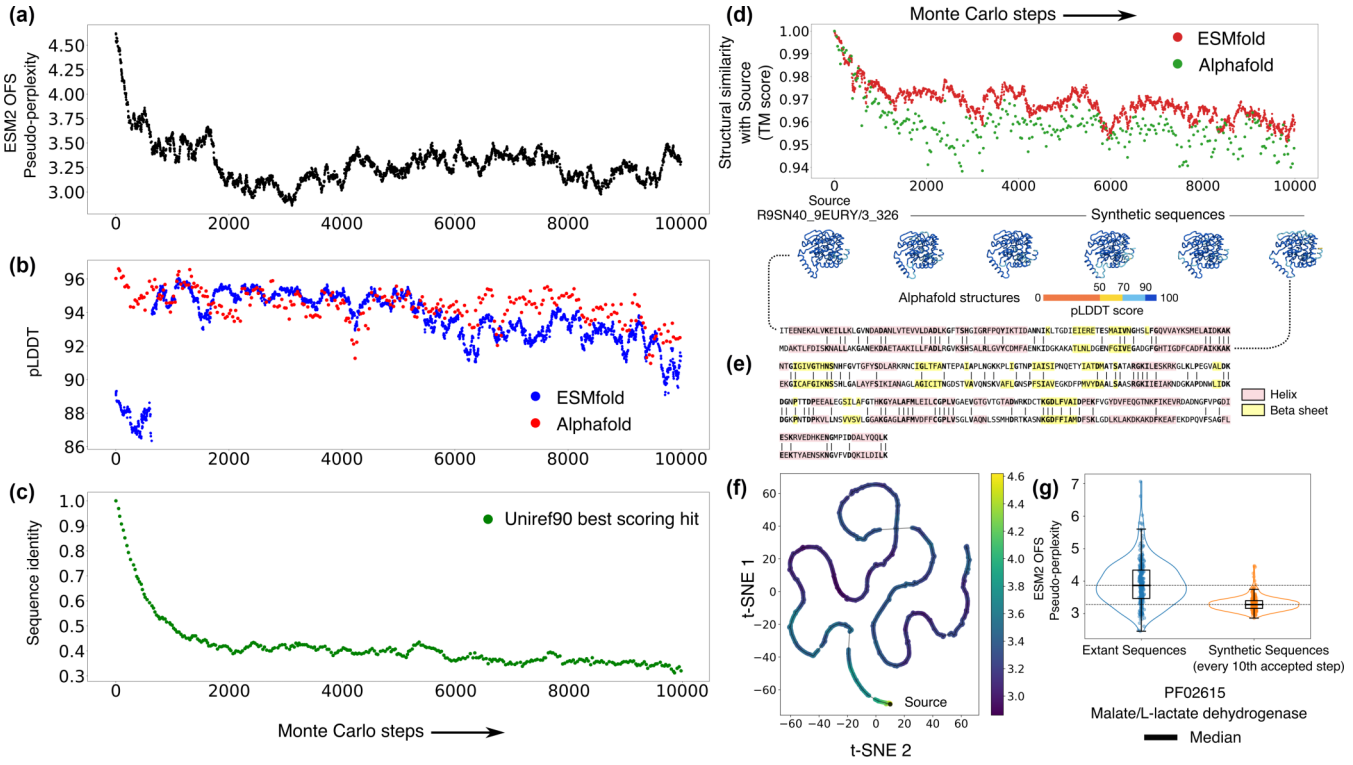

FIG. 4. The OFS masked residue profile and pseudo-perplexity can be used to efficiently explore protein morphospace. (a) A low temperature Monte Carlo run with sequence profile-based mutation proposals and ESM2 OFS pseudo-perplexity as the energy function. (b) The folding confidence of the sequences is consistently high, as measured by the pLDDT score from both Alphafold and ESMFold. (c) The sequence identity that the generated synthetic sequences share with the highest scoring hit in Uniref90 decreases over the course of the procedure. (d) The structural fold of the generated sequences determined via Alphafold and ESMFold remains similar to the fold of the source sequence as measured by the TM score. (e) Structures assumed by the synthetic sequences, folded via Alphafold, along with the sequence alignment between the source sequence and the final synthetic sequence. (f) A visualization of the path traced out by the synthetic sequences through a t-SNE projection of the mean-pooled ESM2 embeddings. The points are colored by their ESM2 OFS pseudo-perplexity. (g) A visualization of the difference between the fitness scores of the two sets of sequences: extant and synthetic in a box plot.

the provided constraints. The generated pool of sequences can be scored to identify the most promising candidates, and the selected candidate sequences can then be experimentally evaluated for downstream use [11,62,67,68].

We contend that the one fell swoop technique combined with an iterative refinement workflow is particularly well suited for the family-specific sequence generation problem. The mutation proposals used in these workflows are typically sampled uniformly over the single residue mutational landscape of the sequence. We propose to use the sequence profile obtained from the ESM2 OFS setup as the sampling distribution for the proposals instead, as it is poised to automatically upsample sensible mutations and therefore increase the exploration efficiency of the procedure. A combination of sequence likelihoods drawn from models trained on different data modalities and modeling objectives could be used to define a custom energy function to generate high fitness variants of the input seed sequence. Such an ensembled setup can potentially lead to higher downstream success by serving as a powerful prefilter for viability and enhanced function in the sequence generation process [45].

As a proof of concept, we use ESM2 OFS pseudo-perplexity as the energy function to generate variants of the Malate/L-lactate dehydrogenase family of enzymes that are predicted to have enhanced stability and function via low-temperature sampling (Fig. 4). The results show that the energy of the generated sequences equilibrates to a lower range of values [Fig. 4(a)] over the course of the run, while the folding confidence of the sequences as measured by the pLDDT score remains high [Fig. 4(b)]. In addition, the generated sequences progressively decrease in their shared sequence identity with sequences in the Uniref90 database [Fig. 4(c)] as determined via mmseqs2 search [26], while the structural similarity with the source structure as measured by the TM score [69] remains high [Fig. 4(d)]. The procedure can thus be used to generate a diverse set of candidate sequences that are predicted to have enhanced properties by enabling explicit control over the fitness of the generated variants through the sampling temperature.

VII. DISCUSSION

Here we introduce one fell swoop, a distillation-inspired trick [70] for fast single-pass estimation of pseudo-perplexity for protein sequences. Distillation involves training a compact model to replicate the output of a larger, more complex model by using the nuanced assessments of the larger model as the target output for the compact model. In a similar vein,

OFS entails training a small projection model to generate the masked profile of a protein sequence, as predicted by the much larger encoding language model. The technique deviates from traditional distillation in allowing the projection model to access the encoder's unmasked residue embeddings to enable accurate and parallelized estimation of the masked profile. The estimated masked profile serves as a valuable intermediate in the pseudo-perplexity calculation as it can be used to identify suboptimally occupied positions in a sequence, and pinpoint the options that fit best in the given context.

This approach circumvents the need to fine-tune the encoder specifically for pseudo-log-likelihood estimation, as has been proposed in the natural language setting [25]. Instead, the specialized OFS model handles the regression task by utilizing the full resolution of the sequence representation from the encoder. This representation consists of the set of all residue embeddings that are generated for each residue within the sequence. This is in contrast to the typical use of a singular coarse-grained representation of the sequence obtained by averaging over the residue embeddings, or the use of the classification token embedding as is done in BERT [14], for sequence-level regression problems of this sort [16,25]. We are leveraging the small size of the vocabulary for protein sequences in making this design choice, as it likely makes the regression task easier.

We used ESM2 (650M) to demonstrate the efficacy of OFS pseudo-perplexity as a fitness estimation metric by benchmarking its performance on the ProteinGym dataset. We show that the scoring scheme enables encoding models to evaluate indels, and more notably, we find that ESM2 OFS pseudo-perplexity defines a new state of the art on the ProteinGym indels benchmark. The metric's strong performance at fitness estimation for stability-based assays prompted us to investigate its potential to detect the elevated stability reported in reconstructed ancestral sequences. We found that reconstructed ancestral sequences tend to have a lower ESM2 OFS pseudo-perplexity relative to extant sequences for the majority of the protein families that we considered.

Reconstructed ancestral sequences thus constitute a class of synthetic sequences that by and large score higher on predicted fitness than natural extant sequences—a trend that is consistent with empirical measurements of protein function. This behavior is noteworthy because the model assigns higher fitness scores to reconstructed sequences when it has only been trained on extant sequences [71]. This feat, much like the model's performance on the fitness estimation benchmark, is a testament to its ability to generalize well beyond what it has seen during training.

We also propose a Markov chain Monte Carlo based iterative refinement workflow for protein design where the sequence profile is used as the proposal distribution, and ESM2 OFS pseudo-perplexity is used as the energy function. The temperature parameter for the process can be tuned to selectively sample sequences that are predicted to have enhanced functional properties. We have observed that short runs reliably generate sequences with a high folding confidence as measured by their pLDDT scores. However, we also found that the folding confidence of the generated sequences eventually degrades over the course of a single very long run. This phenomenon can in part be traced to an effect that arises due to in-context learning, where pseudo-perplexity can remain low even when the sequence being evaluated is nonsensical. The emergence of imperfect repeated motifs within the sequence over the course of an extended run is a hallmark of this effect. In addition, these models have been shown to suffer from phylogenetic biases that stem from uneven species representation in the training dataset. These biases can lead the generated sequences to drift towards favored clades, which can in turn lead to suboptimal design outcomes when working with under-represented sequences [72]. These factors should be carefully considered when using these generative methods. Nevertheless, the model's potential to uncover nonobvious design principles bolsters the case for its use as a vehicle to explore protein morphospace.

ACKNOWLEDGMENTS

We thank Michael Abbott, Jose Betancourt, and Casey Dunn for close reads and useful feedback on the manuscript. This work was supported by NIH R35GM138341 (B.B.M.). We also thank the Yale Center for Research Computing, ITS Cloud, and AWS for guidance and use of AWS Cloud computing infrastructure through Cloud Credits For Research Program.

APPENDIX

1. Pseudo-perplexity in one fell swoop

The one fell swoop setup is composed of an ensemble of eight multilayer perceptron models that use ReLU activations and layer normalization. The input for each MLP is 1280-dimensional followed by layers that progressively reduce the dimensions of the input vector: [1280, 557, 243, 106, 46, 20]. The training data are generated by filtering the Swissprot database to sequences shorter than 700 residues and clustering the remaining sequences at 50% identity via mmseqs2. The representative sequences for the clusters are then used for the sequence profile calculation that is performed by masking the residues one at a time. The probability vector for each position is calculated via the softmax operation over the logits for the 20 natural amino acid residues. Sequences that share greater than 50% identity with reference sequences in the ProteinGym dataset were withheld from being used in training. We use a 99 to 1 training-validation split, and the model is trained to minimize the cross-entropy loss between its prediction and the output from the one-at-a-time calculation via AdamW.

The one hot vector model is trained on the same data splits as the one fell swoop model. We pass the one hot vector representation of protein sequences through a linear layer to upscale it to 1280 dimensions and then feed this upscaled representation to the ensemble of MLP models. The temperature of the softmax operation for the Blosum62 logits is chosen to minimize the loss achieved on the validation split.

The KL divergence between the predictions of two MLP models for a given sequence is calculated by taking the KL divergence between the predictions for each position in the sequence and then taking the mean over these values. In the figure, we show the mean value of the KL divergence between all model pairs, computed over all sequences in the validation split.

2. Evaluating OFS pseudo-perplexity as a fitness estimator

Pseudo-perplexity via the one-at-a-time calculation is calculated for a subset of 61 DMS assays in ProteinGym, where the sum of the lengths of the variants in the assay is less than the chosen threshold of 40 000. The masked marginal heuristic is calculated by masking all mutated positions at the same time, and the probability vector for each position is calculated via the softmax operation over the logits of the 20 natural amino acids. The aggregate mean in Tables I and II is calculated by aggregating over the uniprot id, and then aggregating over all the assays.

3. Ancestral state reconstruction: Old is gold?

a. Decision boundary

ESM2 OFS pseudo-perplexity, used in tandem with Cliff's delta, allows us to measure the difference in fitness between two populations of sequences. This metaphorical instrument has to be calibrated before it can be put to use. This involves characterizing the accuracy of its predictions on data where the ground truth is known, and thereby evaluating its reliability and effectiveness at the task. We generate the evaluation data for this purpose by partitioning the variants in a single DMS assay into two randomly drawn populations. The sampling scheme to partition the sets is tuned by a bias parameter that biases the procedure to selectively sample high or low fitness variants within a given DMS assay. Each partition contains 200 sequences, unless the total number of variants in the assay is fewer than 400. For these cases, we use all the variants in the assay to generate the partitions. Several such partitions are created for all of the 283 assays in ProteinGym yielding diverse sets of populations for which the true value of the disparity in function can be calculated. The true value in this context is defined to be the negative Cliff's delta between the experimentally measured fitness scores of the sequences in the two sets. The negative sign is introduced in the expression to match its sign with the inferred disparity that uses the ESM2 OFS pseudo-perplexity of the sequences in the calculation. The reliability of the reading is assessed by the consistency in the sign of the inferred and true Cliff's delta. This evaluation metric prioritizes getting the direction of the disparity right, even if the degree of the disparity in function between the two sets is not inferred accurately.

The results indicate that the consistency in the sign of the inferred and true value strongly depends on the magnitude of the inferred value: the larger the magnitude, the more likely it is that the sign is inferred correctly. The plots in Fig. 3(a) correspond to a bias parameter of 1 (left), 0.5 (middle), and mixed (right), where the parameter is itself a random number in the mixed case. The plots are generated by creating 1000 sets for each of the 283 assays, and the calculation for the true and inferred Cliff's delta is performed by using the DMS fitness scores and ESM2 OFS pseudo-perplexity, respectively. This allows us to define the range of values of the inferred Cliff's delta that can be used as reliable predictors for the direction of the disparity in function between the two sets. Such a range can take the form of a decision boundary: values that lie within the boundary are discarded as error-prone, and therefore uncertain, while values that lie outside of it can be used as confident estimates. The continuously varying estimate of the disparity in function between sets A and B, measured via Cliff's delta, is thus mapped to three discrete states in this setup: (i) values in A tend to be greater than values in B, (ii) values in A tend to be less than values in B, and (iii) values in the two sets cannot be reliably distinguished by the procedure.

The efficacy of the classification method is predicated on the placement of the boundaries that define the three states. Choosing the precise value for the threshold is a compromise between how much of the true signal we are willing to forego, and how many errors we are willing to tolerate: too stringent of a value for the threshold will inevitably exclude a significant proportion of cases where the discrepancy in function is truly present, whereas a low cutoff will allow errors to creep in. We choose to err on the side of caution and opt for a symmetric threshold of 0.33 around 0. This threshold allows for a high prediction accuracy for the direction of the discrepancy in function, which in turn instills confidence in the method's use as a predictive tool. The higher accuracy comes at the expense of dismissing the readings for a sizable chunk of the pairs in the evaluation sets as unreliable.

We simulate a test run of the procedure on specially constructed populations of sequences for each of the 283 assays: minimal disparity and maximal disparity sets. For assays where the total number of variants is greater than 400, only the top and bottom 200 sequences arranged by the fitness scores are used in this analysis, while all the sequences are used for assays with fewer than 400 variants. The 95% confidence intervals are generated via sampling with repetition.

b. Ancestral state reconstruction

Seed alignments for pfam families with codes from PF00001 to PF03000 were procured from the Interpro database. Only families with a sequence length (including gaps) less than or equal to 500, at least 100 unique sequences and at most 250 total sequences, a gap content of less than 50%, and no nonstandard amino acids were considered for the analysis. These filters yield 301 protein families in total. We used IQTREE2 to infer the phylogeny of the families using the MFP model finder flag. The inferred tree is fed to FastML for ancestral state reconstruction, and the single most likely sequence for each node in the tree is used in the analysis. We ran the outlined procedure for all 301 families; however, FastML jobs only finished for 257 families even after 96-h-long runtimes for the individual jobs. The results for these 257 families are presented in the analysis.

4. Exploring protein morphospace

A state-dependent proposal distribution necessitates a modification in the acceptance probability for the regular Metropolis criterion used in symmetric proposal schemes. We denote the state of the sequence at time t by X_t. Let u be a mutation that changes the sequence X_t to X', and let v be the inverse mutation that changes the sequence X' to X_t. Then the probability of accepting the mutation u for the sequence X_t is

given by

$$A(u|X_t) = \min\left(1, \frac{P(X')}{P(X_t)} \frac{P(v|X')}{P(u|X_t)}\right).$$

Here $P(X)$ is given by $e^{-E(X)/T}$, where E is the energy function and T is the temperature. The probabilities for u and v can be calculated using the sequence profiles of the respective states. We use ESM2 OFS pseudo-perplexity as the energy function in Fig. 4, and we use a temperature of 0.01 for sampling the sequences. Note that only substitutions are used as mutation proposals in Fig. 4.

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with alphafold, Nature (London) **596**, 583 (2021).

[2] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN, Science **378**, 49 (2022).

[3] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles *et al.*, De novo design of protein structure and function with RFdiffusion, Nature (London) **620**, 1089 (2023).

[4] P. Notin, M. Dias, J. Frazer, J. Marchena-Hurtado, A. N. Gomez, D. Marks, and Y. Gal, Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval, in *Proceedings of the 39th International Conference on Machine Learning*, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR, 2022), Vol. 162, pp. 16990–17017.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, Adv. Neural Inf. Proc. Syst. **30** (2017).

[6] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science **379**, 1123 (2023).

[7] E. Nijkamp, J. A. Ruffolo, E. N. Weinstein, N. Naik, and A. Madani, Progen2: exploring the boundaries of protein language models, Cell Syst. **14**, 968 (2023).

[8] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, Msa transformer, in *International Conference on Machine Learning* (PMLR, 2021), pp. 8844–8856.

[9] C. Hsu, R. Verkuil, J. Liu, Z. Lin, B. Hie, T. Sercu, A. Lerer, and A. Rives, Learning inverse folding from millions of predicted structures, in *International Conference on Machine Learning* (PMLR, 2022), pp. 8946–8970.

[10] M. Heinzinger, K. Weissenow, J. G. Sanchez, A. Henkel, M. Steinegger, and B. Rost, Bilingual language model for protein sequence and structure, NAR Genomics and Bioinformatics **6**, lqae150 (2024).

[11] R. Verkuil, O. Kabeli, Y. Du, B. I. M. Wicky, L. F. Milles, J. Dauparas, D. Baker, S. Ovchinnikov, T. Sercu, and A. Rives, Language models generalize beyond natural proteins, bioRxiv 2022–12 (2022).

[12] T. Bepler and B. Berger, Learning the protein language: evolution, structure, and function, Cell Syst. **12**, 654e3 (2021).

[13] J. A. Ruffolo and A. Madani, Designing proteins with language models, Nat. Biotechnol. **42**, 200 (2024).

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (2019), pp. 4171–4186.

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

[16] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger *et al.*, Prottrans: Toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. **44**, 7112 (2021).

[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Improving language understanding by generative pre-training (2018), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

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

[19] N. Ferruz, S. Schmidt, and B. Höcker, Protgpt2 is a deep unsupervised language model for protein design, Nat. Commun. **13**, 4348 (2022).

[20] D. Hesslow, N. Zanichelli, P. Notin, I. Poli, and D. Marks, Rita: A study on scaling up generative protein sequence models, arXiv:2205.05789.

[21] J. Horne and D. Shukla, Recent advances in machine learning variant effect prediction tools for protein engineering, Indust. Eng. Chem. Res. **61**, 6235 (2022).

[22] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, pp. 29287–29303.

[23] P. Notin, A. Kollasch, D. Ritter, L. Van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Orenbuch, R. Weitzman *et al.*, Proteingym: Large-scale benchmarks for protein fitness prediction and design, Adv. Neural Inf. Proc. Syst. **36** (2024).

[24] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher *et al.*, Large language models generate functional pro-

tein sequences across diverse families, Nat. Biotechnol. **41**, 1099 (2023).

[25] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, Masked language model scoring, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (ACL, 2020), pp. 2699–2712.

[26] M. Steinegger and J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, Nat. Biotechnol. **35**, 1026 (2017).

[27] D. M. Fowler and S. Fields, Deep mutational scanning: A new style of protein science, Nat. Methods **11**, 801 (2014).

[28] J. Su, C. Han, Y. Zhou, J. Shan, X. Zhou, and F. Yuan, SaProt: protein language modeling with structure-aware vocabulary (2024), https://openreview.net/forum?id=6MRm3G4NiU.

[29] M. Li, Y. Tan, X. Ma, B. Zhong, H. Yu, Z. Zhou, W. Ouyang, B. Zhou, P. Tan, and L. Hong, Prosst: Protein language modeling with quantized structure and disentangled attention, Advances in Neural Information Processing Systems **37**, 35700 (2024).

[30] S. Paul, A. Kollasch, P. Notin, and D. Marks, Combining structure and sequence for superior fitness prediction, in *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop* (2023).

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

[32] P. Kapli, Z. Yang, and M. J. Telford, Phylogenetic tree building in the genomic age, Nat. Rev. Genet. **21**, 428 (2020).

[33] J. B. Joy, R. H. Liang, R. M. McCloskey, T. Nguyen, and A. F. Y. Poon, Ancestral reconstruction, PLoS Comput. Biol. **12**, e1004763 (2016).

[34] G. K. A. Hochberg and J. W. Thornton, Reconstructing ancient proteins to understand the causes of structure and function, Annu. Rev. Biophys. **46**, 247 (2017).

[35] A. S. Pillai, S. A. Chandler, Y. Liu, A. V. Signore, C. R. Cortez-Romero, J. L. P. Benesch, A. Laganowsky, J. F. Storz, G. K. A. Hochberg, and J. W. Thornton, Origin of complexity in haemoglobin evolution, Nature (London) **581**, 480 (2020).

[36] S. F. Field, M. Y. Bulina, I. V. Kelmanson, J. P. Bielawski, and M. V. Matz, Adaptive evolution of multicolored fluorescent proteins in reef-building corals, J. Mol. Evol. **62**, 332 (2006).

[37] V. J. Lynch, G. May, and G. P. Wagner, Regulatory evolution through divergence of a phosphoswitch in the transcription factor cebpb, Nature (London) **480**, 383 (2011).

[38] M. C. Nnamani, S. Ganguly, E. M. Erkenbrack, V. J. Lynch, L. S. Mizoue, Y. Tong, H. L. Darling, M. Fuxreiter, J. Meiler, and G. P. Wagner, A derived allosteric switch underlies the evolution of conditional cooperativity between hoxa11 and foxo1, Cell Rep. **15**, 2097 (2016).

[39] K. J. Brayer, V. J. Lynch, and G. P. Wagner, Evolution of a derived protein–protein interaction between hoxa11 and foxo1a in mammals caused by changes in intramolecular regulation, Proc. Natl. Acad. Sci. USA **108**, E414 (2011).

[40] N. M. Hendrikse, A. Sandegren, T. Andersson, J. Blomqvist, Å. Makower, D. Possner, C. Su, N. Thalén, A. Tjernberg, U. Westermark *et al.*, Ancestral lysosomal enzymes with increased activity harbor therapeutic potential for treatment of hunter syndrome, iScience **24**, 102154 (2021).

[41] E. A. Gaucher, S. Govindarajan, and O. K. Ganesh, Palaeotemperature trend for precambrian life inferred from resurrected proteins, Nature (London) **451**, 704 (2008).

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

[43] V. A. Risso, J. A. Gavira Gallardo, J. M. Sánchez Ruiz *et al.*, Thermostable and promiscuous precambrian proteins (2014).

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

[45] S. R. Johnson, X. Fu, S. Viknander, C. Goldin, S. Monaco, A. Zelezniak, and K. K. Yang, Computational scoring and experimental evaluation of enzymes generated by neural networks, Nat. Biotechnol. **43**, 396 (2025).

[46] A. L. Pey, D. Rodriguez-Larrea, S. Bomke, S. Dammers, R. Godoy-Ruiz, M. M. Garcia-Mira, and J. M. Sanchez-Ruiz, Engineering proteins with tunable thermodynamic and kinetic stabilities, Proteins: Struct. Funct. Bioinform. **71**, 165 (2008).

[47] B. J. Sullivan, T. Nguyen, V. Durani, D. Mathur, S. Rojas, M. Thomas, T. Syu, and T. J. Magliery, Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability, J. Mol. Biol. **420**, 384 (2012).

[48] R. S. Komor, P. A. Romero, C. B. Xie, and F. H. Arnold, Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods, Protein Eng. Des. Select. **25**, 827 (2012).

[49] M. Sternke, K. W. Tripp, and D. Barrick, Consensus sequence design as a general strategy to create hyperstable, biologically active proteins, Proc. Natl. Acad. Sci. USA **116**, 11275 (2019).

[50] M. Lynch and A. Force, The probability of duplicate gene preservation by subfunctionalization, Genetics **154**, 459 (2000).

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

[52] J. Huerta-Cepas, F. Serra, and P. Bork, ETE 3: reconstruction, analysis, and visualization of phylogenomic data, Mol. Biol. Evol. **33**, 1635 (2016).

[53] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski *et al.*, Biopython: Freely available python tools for computational molecular biology and bioinformatics, Bioinformatics **25**, 1422 (2009).

[54] N. Cliff, Dominance statistics: Ordinal analyses to answer ordinal questions. Psych. Bull. **114**, 494 (1993).

[55] T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell *et al.*, InterPro in 2022, Nucl. Acids Res. **51**, D418 (2023).

[56] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson *et al.*, Pfam: The protein families database in 2021, Nucl. Acids Res. **49**, D412 (2021).

[57] B. Q. Minh, M. W. Hahn, and R. Lanfear, New methods to calculate concordance factors for phylogenomic datasets, Mol. Biol. Evol. **37**, 2727 (2020).

[58] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin, ModelFinder: Fast model selection

for accurate phylogenetic estimates, Nat. Methods **14**, 587 (2017).

[59] H. Ashkenazy, O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, and T. Pupko, FastML: A web server for probabilistic reconstruction of ancestral sequences, Nucl. Acids Res. **40**, W580 (2012).

[60] S. Arora and A. Goyal, A theory for emergence of complex skills in language models, arXiv:2307.15936.

[61] P. Notin, N. Rollins, Y. Gal, C. Sander, and D. Marks, Machine learning for functional protein design, Nat. Biotechnol. **42**, 216 (2024).

[62] B. Hie, S. Candido, Z. Lin, O. Kabeli, R. Rao, N. Smetanin, T. Sercu, and A. Rives, A high-level programming language for generative protein design, bioRxiv 2022–12 (2022).

[63] K. H. Sumida, R. Núñez-Franco, I. Kalvet, S. J. Pellock, B. I. M. Wicky, L. F. Milles, J. Dauparas, J. Wang, Y. Kipnis, N. Jameson *et al.*, Improving protein expression, stability, and function with ProteinMPNN, J. Am. Chem. Soc. **146**, 2054 (2024).

[64] R. W. Shuai, J. A. Ruffolo, and J. J. Gray, Iglm: Infilling language modeling for antibody sequence design, Cell Syst. **14**, 979 (2023).

[65] T. Truong Jr. and T. Bepler, PoET: A generative model of protein families as sequences-of-sequences, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2023), Vol. 36, pp. 77379–77415.

[66] D. Sgarbossa, U. Lupo, and A.-F. Bitbol, Generative power of a protein language model trained on multiple sequence alignments, Elife **12**, e79854 (2023).

[67] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church, Low-n protein engineering with data-efficient deep learning, Nat. Methods **18**, 389 (2021).

[68] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera *et al.*, De novo protein design by deep network hallucination, Nature (London) **600**, 547 (2021).

[69] C. Zhang, M. Shine, A. M. Pyle, and Y. Zhang, Us-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes, Nat. Methods **19**, 1109 (2022).

[70] G. Hinton, O. Vinyals, and J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531.

[71] A. Shaw, H. Spinner, J. Shin, S. Gurev, N. Rollins, and D. Marks, Removing bias in sequence models of protein fitness, bioRxiv 2023–09 (2023).

[72] F. Ding and J. Steinhardt, Protein language models are biased by unequal sequence sampling across the tree of life, *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design* (2024).