Perspective

Transforming molecular evolutionary analyses with protein foundation models

Sudhir Kumar[1,2]*, Rohan Alibutud[1,2], John Allard[1,2], Mohammad E. Mowlaei[1,3], and Xinghua Shi[3]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA.

[2]Department of Biology, Temple University, Philadelphia, PA 19122, USA.

[3]Department of Computer and Information Science, Temple University, Philadelphia, PA 19122, USA.

Abstract

Transformer-based protein foundation models (pFMs) are emerging as powerful models for analyzing protein sequences. However, broader adoption of pFMs in molecular evolutionary analysis requires articulation of their evolutionary foundations and how they relate to traditional models and workflows. We synthesize research showing that pFMs trained on millions of natural protein sequences can reveal residue-residue dependencies, recover evolutionary constraints, predict pathogenic mutations, and reconstruct sequence relationships. We introduce the main outputs of pFM-based analysis and demonstrate that they reflect neutral evolutionary principles without relying on explicit amino acid substitution matrices, phylogenies, or even multiple sequence alignments. We suggest that single-sequence analysis with pFMs will streamline many molecular evolutionary analyses and expand their reach to proteins with few or no known homologs. As a result, pFMs will complement and extend traditional approaches, advancing the field from phenomenological models of primary-sequence differences to models of residue dependencies in proteins. We also discuss limitations, computational considerations, and open questions regarding the routine use of transformer models in molecular evolutionary research.

A. Introduction

High-throughput molecular sequencing has revealed a vast reservoir of protein diversity spanning the tree of life, providing unprecedented opportunities to understand how protein structures and functions evolve.[1–6] Classical molecular evolutionary analyses typically begin by assembling a multiple sequence alignment (MSA) and then applying an amino acid substitution model within a statistical framework (**Fig. 1a**). These approaches have been successful in understanding the tempo and mode of protein evolution and diversification across many scientific disciplines and in thousands of studies[7] each year.

Traditional amino acid substitution models have been developed using several simplifying assumptions. Protein sites are assumed to evolve independently, despite long-recognized patterns of coevolution driven by structural contacts and functional couplings.[8–11] A single substitution pattern is applied across all protein sites and evolutionary lineages to generate sufficient counts of evolutionary substitutions for statistical analysis, thereby imposing the assumptions of stationarity, reversibility, and homogeneity of substitution models across sites and species within MSAs.[12,13] The primary objectives of these analyses are to estimate relative rates of residue substitutions, branch lengths in phylogenies, and other evolutionary parameters that best describe the observed sequence differences in the MSA. In most applications, sites are treated as independent and identically distributed, so the influences of local and long-range sequence context on constraint are only partially captured, for example, through discrete or continuous rate heterogeneity models.[13]

Transformer-based deep learning[15] of molecular evolutionary inference offers a contrasting modeling approach that focuses on residue dependencies in proteins arising from epistasis and structural coupling.[10,14,16,17] Transformers are neural networks with many layers and nodes, each with parameters (weights), optimized (learned) via masked language modeling, a technique originally developed for natural language processing.[15] In masked modeling, a fraction of amino acid residues is randomly selected and replaced with a special token (<mask>). The neural network then predicts the identities of these hidden residues by exploiting statistical dependencies among the unmasked residues and the network's internal representations and weights (parameters). These weights are iteratively adjusted to maximize the likelihood of correct residue recovery, implemented through loss functions such as categorical cross-entropy.[15] These loss functions have Bayesian interpretations, where the model’s predictions correspond to posterior probabilities over residues given the observed context.[18]

Masked modeling is a form of self-supervised learning in which millions to billions of weights are estimated, yielding a protein foundation model (pFM) without relying on explicit stochastic models, parameterized substitution matrices, or assumptions of time-reversibility and stationarity that are required to estimate parameters in traditional evolutionary analysis (pFMs are also referred to as protein language models). Still, the objective of optimizing weights (*aka* learning) to build a pFM parallels the use of statistical frameworks in conventional evolutionary methods. Both estimate model parameters: the former seeks optimal residue substitution rates and branch lengths in a phylogenetic context, and the latter tunes millions of network weights using the conditional probabilities of residues in sequence space. It has been argued that mask modeling is analogous

to high-dimensional likelihood-based modeling, as it makes the observed data most probable under the network model.[18,19]

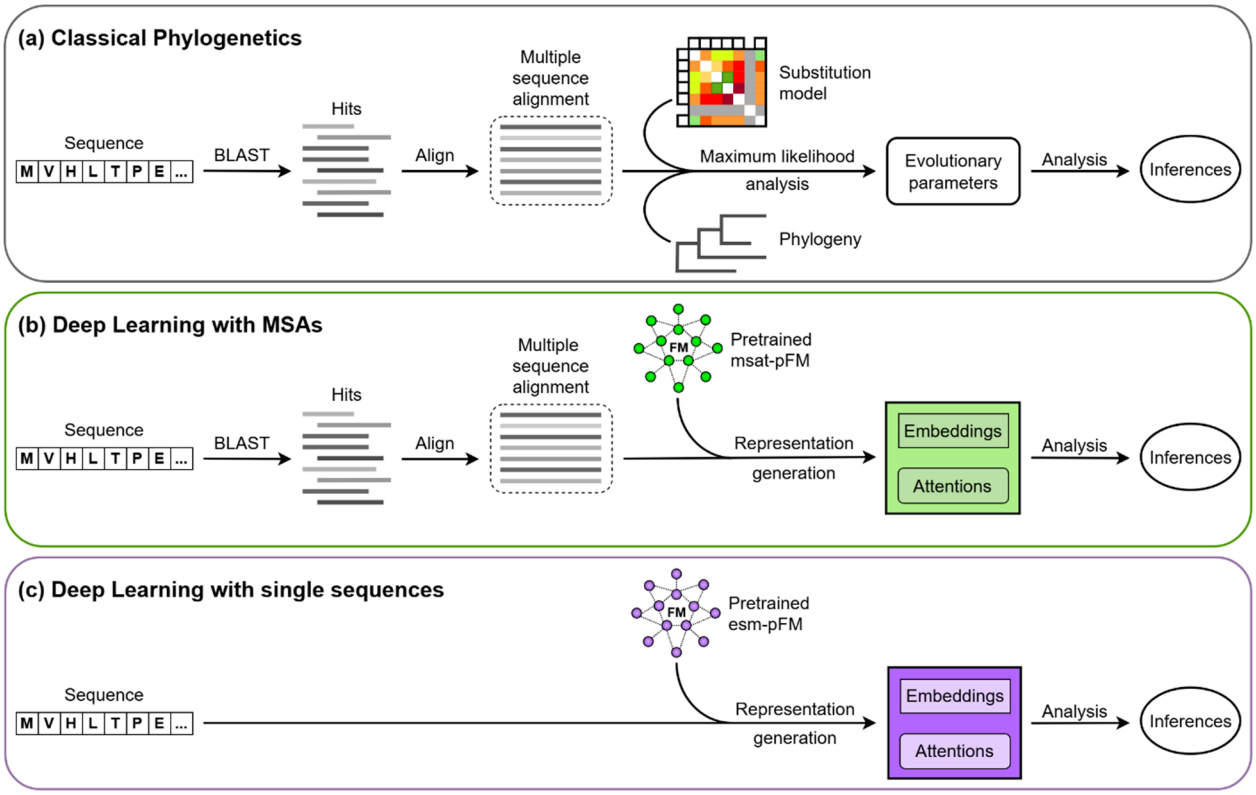


Figure 1. Workflow for molecular evolutionary analysis using classical and deep learning models. (**a**) Traditional phylogenetic analysis of protein sequences starts with the assembly of multiple sequence alignments (MSAs) that are analyzed statistically with amino acid substitution models (e.g., **Fig. 2**) and optimality principles, such as maximum likelihood, to produce evolutionary insights. (**b**) The deep learning analysis using an MSA transformer model[14] differs in that a pretrained protein foundation model (msat-pFM, **Fig. 3**) is used, after an MSA is assembled, to estimate dependencies, residue embeddings, and attentions via Representation Generation (RepGen, see **Box 2**) that does not need to involve statistical principles or optimality criteria. (**c**) In contrast, a protein sequence can be subjected to deep learning analysis using a protein foundation model pretrained on unaligned individual protein sequences via the Evolutionary Scale modeling (ESM)[10] approach (esm-pFM, **Fig. 5**), with Representation Generation used for downstream inference. This workflow differs from the other two in that it bypasses MSA construction and instead directly produces residue embeddings and attentions using a pFM.

Protein FMs and traditional models are complementary ways of describing biology. Classical substitution models are phenomenological, as they describe how often residues change along branches of a tree given a particular alignment, and are optimized for tasks such as phylogeny inference and rate estimation. Protein FMs are trained to predict residues from their sequence context across the universe of proteomes, so their parameters are more closely tied to the compatibility of residue configurations that folding, function, and cellular quality control will tolerate. In this sense, pFMs approximate a more functional, mechanistic constraint structure rooted in residue-residue and higher-order interactions. In contrast, traditional stochastic modeling focuses on summarizing the historical outcomes of those constraints. This

complementarity means that the use of pFMs in molecular evolutionary analysis represents a significant conceptual shift, in which residue interactions evolve and define the permissible space of residue types at each position in every protein, rather than as a set of independent residue-substitution events.

Protein FMs typically have billions of parameters, estimated from a large and diverse corpus of protein sequences, without an explicit substitution-rate matrix or statistical distributions. Instead, protein FMs internalize high-dimensional patterns of residue dependencies within a neural network architecture. Their parameters are distributed across network layers that implement context-dependent functions. For each residue, the model aggregates information from multiple positions and determines which amino acids are compatible in that context. Because there are many possible configurations of pairwise and higher-order residue dependencies, the effective parameter space is enormous. This is unsurprising because statistical models of epistasis, such as the Potts model, can have millions of parameters, requiring thousands of aligned sequences for statistical estimation.[20] This is why pFMs require both large capacity and massive collections of protein sequences.

Conversely, traditional comparative sequence analysis methods estimate only a fraction of the parameters in pFMs, such as substitution rates between residues, branch lengths on a phylogeny, and parameters describing rate variation among sites.[12,13] These parameters are estimated from a single multiple-sequence alignment; they are fitted to account for the observed pattern of differences among a limited number of homologous sequences and sites. The result is a phenomenological description of past evolution in the given MSA: a summary of which substitutions occur more frequently, often with biological correlates, such as the well-known correspondence between PAM values and physicochemical distance measures.[21]

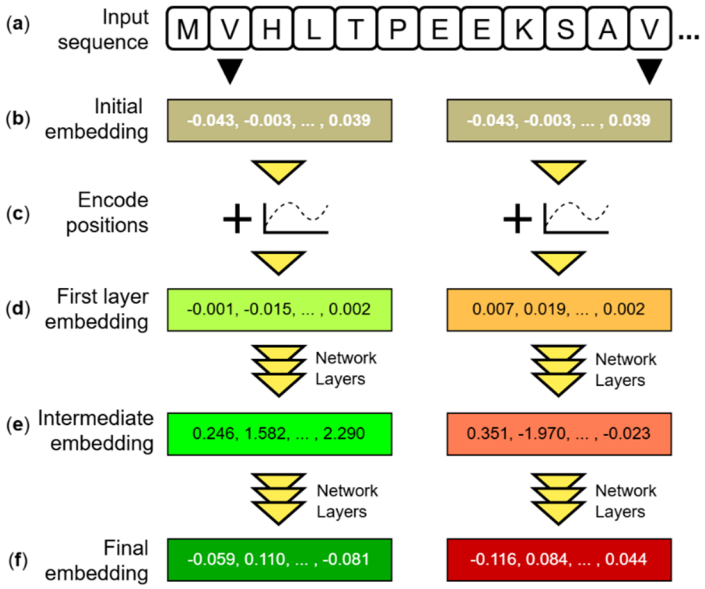
Traditional and deep learning approaches differ in the way they analyze the sequence or alignment of interest. Analyses using deep learning models do not employ commonly used statistical techniques in molecular phylogenetics, such as maximum likelihood, to estimate phylogenetic branch lengths, substitution rate parameters, or optimality principles. They use the pretrained pFM to generate representations (**Fig. 1b** and **c**), which we refer to as the Representation Generation (RepGen) procedure that does not involve statistical optimization or network retraining. RepGen is a set of computational steps in which each neural network layer generates a numerical representation (embeddings) of each residue, a high-dimensional vector (e.g., 520 - 1280 elements) that encodes its context in the sequence and likely residue identity (**Box 1**).

For example, the initial embedding of valine (V) at the second and twelfth sites in the human hemoglobin β sequence is the same, but they diverge as the protein traverses the network layers in **Box 1**. The final embeddings of a residue encode its relationships with all other residues in the sequence, in addition to its identity and position, and are informed by the pretrained pFM. In addition, in each layer of the model, information may be exchanged between the embeddings for residues via the so-called attention mechanism,[22] which produces a matrix of pairwise “attention scores” that govern the extent to which information may be exchanged between pairs of residues. In each layer, multiple instances of the attention mechanism, referred to as attention heads, operate in parallel, each facilitating the mixing of information in different ways that emerged during

model pretraining. Thus, RepGen produces residue embeddings and attention maps that have been used to make evolutionary and functional inferences (**Figs 3** and **5**).

BOX 1 | Representation Generation (RepGen) using a pretrained pFM

Representation Generation (RepGen) is referred to in the literature as a forward pass or inference. RepGen is a sequence of computational steps that converts an amino-acid sequence into numerical representations (embeddings) using a pretrained pFM. No learning or optimization occurs; the model simply applies its context-informed weights in a single forward pass. The first step is to convert each residue in the input protein (e.g., the human hemoglobin β chain a) into a high-dimensional numerical vector, called an embedding. These residue embeddings are learned during FM pretraining. All residues of the same type (e.g., Valine, V) start with the same initial embedding (**b**). Depending on the pFM design, positional embeddings are commonly applied either after the initial embedding or within each layer (**c**). Consequently, each sequence embedding encodes the residue identities and positions in the protein (**d**). The reside embeddings are updated by passing through the pFM network layers (**e**). Each layer combines information from all other residues, producing updated embeddings that capture the extent to which each residue is influenced by others. The final embedding (**f**) for each residue encodes its modeled relationships to all other residues in the sequence, in addition to its identity and position. These embeddings, together with attention maps, form the basic outputs of RepGen that we discuss throughout this Perspective.



This Perspective situates pFMs within the historical development of evolutionary analysis, from early global substitution matrices to coevolutionary statistical models, and highlights how the advent of pFMs extends and complements traditional approaches for studying protein evolution. Our objective is to demystify the use of deep learning for biologists who are accustomed to comparative analysis of protein sequences using traditional models and methods in molecular evolution and phylogenetics. We draw parallels between the behavior of RepGen with pretrained FMs and the expectations of the neutral theory of molecular evolution[23,24]. We suggest that FMs can complement traditional statistical models by enabling single-sequence evolutionary analysis and advance the field from phenomenological modeling of residue differences to the study of the evolution of residue interactions.

This perspective focuses exclusively on pFMs because the principles highlighted here, such as mechanistic constraints, epistasis, and context-dependent compatibility, have been explored in greater detail in protein models. Nucleotide and general biological FMs[25–27] are a promising frontier and fall outside the primary scope of this perspective.

B. Early Foundation-like Models for Protein Evolution

In the late 1970s, the Point Accepted Mutation (PAM) model was published, which described the rate at which a residue type replaces another over evolutionary time.[21] In the iconic 20×20 PAM matrix (**Fig. 2d**), diagonal elements quantified residue conservation, such as the exceptional stability of cysteine (C) and tryptophan (W), while off-diagonal entries reflected specific substitution probabilities, such as the frequent exchange of aspartic acid for glutamic acid. PAM is a global model of residue change, arguably the field's first foundational substitution model, because it was inferred by aggregating phylogenetically inferred residue substitutions across protein families from a modest set of relevant alignments available at that time (**Fig. 2a**).

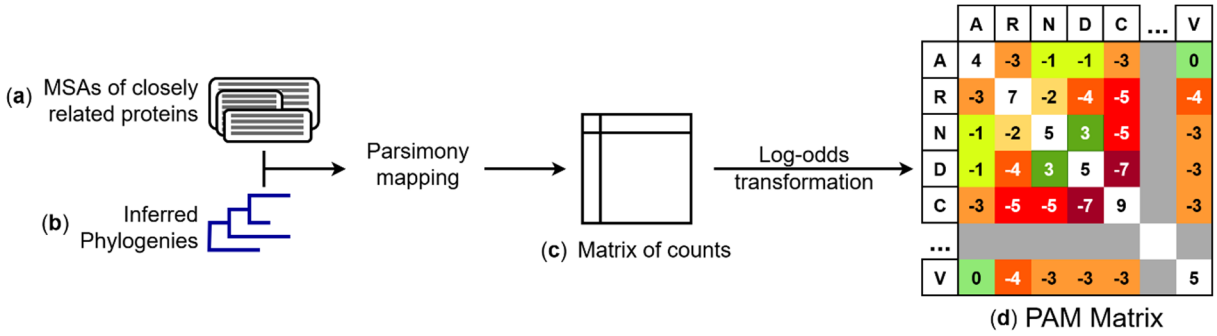


Figure 2. The Point Accepted Mutation (PAM) model for protein evolution.
(**a**) Originally derived from known protein sequences in 1978, the Point Accepted Mutation (PAM)[21] model presents the substitution probability between residues. The number of accepted point mutations is empirically derived from counts of residue substitutions (**c**) inferred from mapping multiple sequence alignments (MSAs) onto phylogenies (**b**) for a set of closely related proteins. (**d**) 20×20 PAM matrix showing log-odds of substitution probabilities with backgrounds in green (evolutionarily common substitutions) to red (evolutionarily uncommon substitutions), with each row and column showing a particular amino acid residue.

PAM can be scaled across time depths (e.g., PAM1 and PAM250) and adapted through empirical frequency (F) corrections for individual proteins (e.g., PAM+F). Its inference was task-agnostic, similar to that of FMs trained via self-supervised learning. However, PAM models have served as a statistical lens for examining protein evolution for decades, becoming indispensable for phylogenetic reconstruction, sequence alignment, functional inference, and pathogenicity annotations.[28–30] PAM also inspired many successors, such as JTT[31] and BLOSUM[32], which refined residue-substitution matrices for specialized datasets and employed more sophisticated statistical methods. These PAM-family models are also generative, in that they have been used to simulate plausible evolutionary trajectories from a starting sequence and a phylogeny.[21]

PAM-family models assume independence of evolutionary substitutions across sites, as well as homogeneity, stationarity, and reversibility of evolutionary substitutions, because substitutions inferred from site-by-site analyses are aggregated across sequences in alignments that represent diverse evolutionary lineages and proteins (**Fig. 2** and **3a**). Statistical methods can relax some of these assumptions for individual proteins and domains[33,34], but the paucity of sufficient substitutions to reliably estimate parameters in standard statistical frameworks is a stumbling

block when analyses are restricted to particular proteins, domains, and/or lineages. Beyond such global single-site models, a variety of models have been developed that allow substitution rates to depend on local sequence features[35], partially relaxing the assumption of independent and identically distributed sites. These extensions can realistically capture short-range contextual effects. However, their state spaces grow rapidly with context size, so they are typically restricted to small motifs and still fall short of the many-body residue interactions that shape protein structures and functions.

Before the emergence of transformer-based protein FMs, a major advance in modeling residue dependencies was the Potts models and their deep-learning extensions.[20,36] These approaches infer pairwise (and sometimes higher-order) couplings between residues by fitting statistical energy landscapes to MSAs containing many sequences. Potts models have demonstrated that coevolutionary signals encode residue-residue contacts, functional constraints, and epistatic interactions, thereby extending beyond classical substitution models that assumed site independence. DeepSequence[37] and EVE[38] extended this idea by using variational autoencoders, enabling generative modeling of sequence families and improving variant-effect prediction.

However, these methods require MSAs with thousands of homologs and depend on explicit probabilistic formulations tied to a specific domain family. Their parameter count scales quadratically with the sequence length, which limits practical applications to relatively short, well-sampled domain families. In contrast, pFMs learn residue dependencies across millions of heterogeneous sequences with or without alignment, allowing them to generalize epistasis and constraints without domain-specific training and to operate even when alignments are sparse or unavailable. Thus, the historical progression from global substitution matrices to Potts models and then to transformers reflects a steady relaxation of independence assumptions and an expansion in representational capacity, even though each step is built on a distinct conceptual and computational framework.

C. Foundation models for sequence alignments

Fifty years after the introduction of the PAM model, a deep learning protein foundation model (pFM) was estimated using a neural network with a transformer architecture (**Fig. 3**)[40]. The MSA-Transformer (msat) modeling employed 26 million sequence alignments, with an average alignment depth of 1192 sequences, from UniProt, without incorporating information on species of origin, molecular function, or phylogeny.[40] We refer to the resulting foundation model as msat-pFM, which has 100 million parameters distributed across 144 attention heads in 12 layers.[40] The weights of these many parameters correspond to patterns of residue co-occurrence both within and across aligned sequences. For downstream inference, the same pretrained msat-pFM can be used for each new alignment to perform inference and generate informative data representations via representation generation (RepGen, **Fig. 1b** and **3**). This contrasts sharply with classical phylogenetic analyses, where a global (or data-specific) statistical model is used under a statistical framework, such as Maximum Likelihood (**Fig. 1a**). In contrast, RepGen analysis is a set of specific calculations in which the input data are passed through all the network layers to produce residue embeddings and attention scores (**Box 1**). No optimization or model building is involved at this stage.

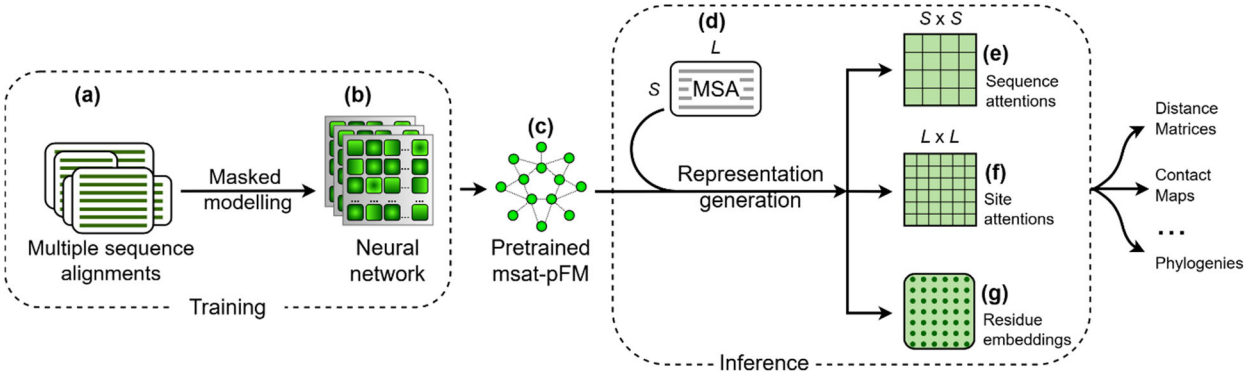


Figure 3. Training and applying a protein foundation model based on sequence alignments.
(**a**) An extensive database of multiple sequence alignments (MSAs) is used to train a transformer-based neural network. (**b**) The MSA-transformer[14] is shown, comprising 12 layers, each with 12 attention heads, represented as squares. Network weights (parameters) are optimized via self-supervised training with masked modeling[15]. (**c**) The training produces a protein foundation model (pFM) that has learned residue dependencies from the input MSA collection. (**d**) The pretrained network (msat-pFM) can be used to analyze any input MSA by Representation Generation (RepGen, **Box 1**), which produces (**e**) sequence attention matrices of dimensions *S x S* where *S* is the number of sequences in the alignment, (**f**) site attention matrices of dimensions *L x L* where *L* is the length of each sequence in the alignment, and (**g**) residue embeddings, which are of variable dimensions depending on the model used. These outputs can be used to make evolutionary inferences, such as estimating evolutionary distances[17], classifying tumor cells[39], and annotating pathogenic variants.[14]

Attentions across sequences at a site. RepGen analysis in pFMs produces attention scores at each sequence site, capturing the relationships among residues.[15] For an alignment with *S* sequences and *L* sites, every attention head generates *L* matrices of size *S* × *S* (**Fig. 3f**). Each cell in these matrices can be interpreted as a context-dependent similarity score. If sequence A pays high attention to sequence B at a site, this may be construed to mean that the residue in B provides strong statistical support for recovering the masked residue in A at that position. **Figure 4b-c** illustrates attentions for sites in an MSA of hemoglobin β (HBB, **Fig. 4a**) with regard to human, in chimpanzee, mouse, and chicken, respectively. At site #7, the chimpanzee sequence has the same residue as the human sequence (E), unlike the mouse (A), which correlates with greater attention from the chimpanzee than from the mouse (**Fig. 4b**). Even though the chicken protein has the same residue (E) as the human protein, its attention is lower, presumably because the residues neighboring E are quite different between human and chicken as compared to human and mouse proteins. This exemplifies pFM's context awareness via the implemented attention mechanism.

At site #53, chimpanzees again show high attention to humans, due to identical residue and context, as compared to mice and chickens both of which show a different residue and have multiple contextual differences from the human sequence (**Fig. 4c**). These patterns of relative attention scores are consistent with humans' more recent divergence times from chimpanzees than mice or chickens (**Fig. 6b**). From these examples, chosen to display the interplay of attentions and context, it is clear that sequence attentions do not merely reflect exact residue identity, but can capture contextual interchangeability where the same residue found in different sequence backgrounds produces different results.

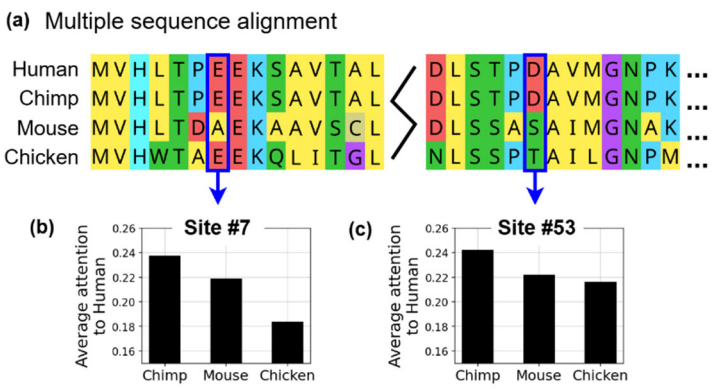


Figure 4. Analysis of a sequence alignment using a foundation model. (**a**) A multiple sequence alignment of the hemoglobin β chain proteins (HBB) in four species: Human, Chimpanzee (Chimp), Mouse, and Chicken. Average sequence attentions across all heads for humans at site #7 (**b**), which harbors the mutation that causes sickle cell anemia, and at site #53 (c). The background is shaded in colors corresponding to biochemical properties of amino acid types.

Averaged sequence attention matrices. The above example shows that pFM analysis can encode evolutionary similarity, which is more flexible than strict sequence matching. We can derive a single matrix of pairwise sequence attentions for each attention head, averaged across all sites in the alignment. This procedure yields 144 matrices, which have been used to train specialized models to produce classical evolutionary distances[17,41], demonstrating that attentions in pFMs can capture evolutionary information. Furthermore, all neighbor-joining (NJ[42]) trees inferred from pairwise distance matrices estimated directly from attention matrices[43] produced phylogenies consistent with the known evolutionary relationships among these four species. Similar successes using attention-derived distance matrices have been reported for longer alignments with more sequences.[43]

An emerging frontier in pFMs is the development of methods that leverage sequence attention scores to reconstruct evolutionary relationships.[43,44] They are attractive because traditional phylogenetic methods must make many simplifying assumptions, such as independence of substitutions across sites, as well as stationarity, reversibility, and homogeneity of the amino acid substitution process. Such assumptions are expected to be frequently violated in conventional phylogenies, particularly for highly diverged sequences. However, it remains unclear whether the use of sequence attentions and their derivatives can outperform traditional methods, as analysis with pFMs does not require making many explicit assumptions. Nonetheless, it is worth noting that many machine learning methods have been proposed for direct phylogeny inference[45,46], which typically require explicit alignments or simulated training data and address a fundamentally different problem than the one considered here, so we have not discussed them here.

Site attention matrices. While sequence attentions may reflect evolutionary relationships among homologous sequences, analysis with pFMs captures another layer of dependency: how residues within a protein co-vary across the alignment positions. These within-sequence dependencies are encoded as site attentions, which illuminate the structural and functional couplings that shape protein evolution. Each attention head tending to protein sequences of length L generates an $L \times L$ site attention matrix because site attentions are shared across sequences during FM pretraining (**Fig. 3e**). Pairwise site attention values are conceptually similar to coupling scores derived from Potts models[17,47]. A key distinction is that Potts model inference typically requires alignments containing thousands of homologous sequences [48,49], whereas site attentions can be estimated for virtually any alignment using a pre-trained FM. While Potts models rely on pairwise coupling

terms, pFMs can represent higher-order and long-range dependencies spanning multiple residues through their attention mechanisms.[9,40] In general, site attention maps have proven remarkably effective in identifying residue-residue contacts, predicting protein folds, and highlighting functionally important sites, without any structural or functional supervision during training.[17,50,51]

D. Protein FMs for raw protein sequences

Evolutionary Scale Modeling (ESM)[10] trains a pFM using raw, unaligned sequences. Each sequence is input independently, and the model learns statistical patterns in residue usage and dependency from hundreds of millions of sequences drawn from databases such as UniParc.[10] Here, we refer to this category of model as esm-pFM (**Fig. 5c**). The first example of an esm-pFM, ESM-1b[10], was developed before msat-pFM because ESM was an extension of modelling designed initially for natural language processing.[10,15] Unlike msat-pFM, which shares residue attentions across homologous sequences within an alignment, esm-pFM learns patterns of residue dependencies from individual sequences without involving alignments. During training, the network encounters residue patterns that interact in folded structures and learns to recognize them because the coevolution of interacting residues produces informative signals for predicting residue identity[10,52] during masked modeling. Consequently, esm-pFMs can directly capture coevolutionary couplings, revealing residue contact maps from single-protein inputs. It is a member of a broader family of pFMs that yield useful embeddings for predicting diverse functional and structural properties across large protein spaces, underscoring the richness of the learned representations.[50,52,53]

For inference using an esm-pFM, a protein sequence of length *L* is passed through the pretrained model (**Fig. 5d**). Then, *L*×*L* matrices of site attentions can be extracted from each attention head. For example, the ESM-2 model, which consists of 33 layers with 20 attention heads each (**Fig. 5b**), yields 660 site attention matrices for the input sequence (**Fig. 5e**, **f**). Such site attention matrices have been used to predict coevolving residues in proteins[52]. Because esm-pFM is for individual sequences, it does not generate any sequence attention matrices.

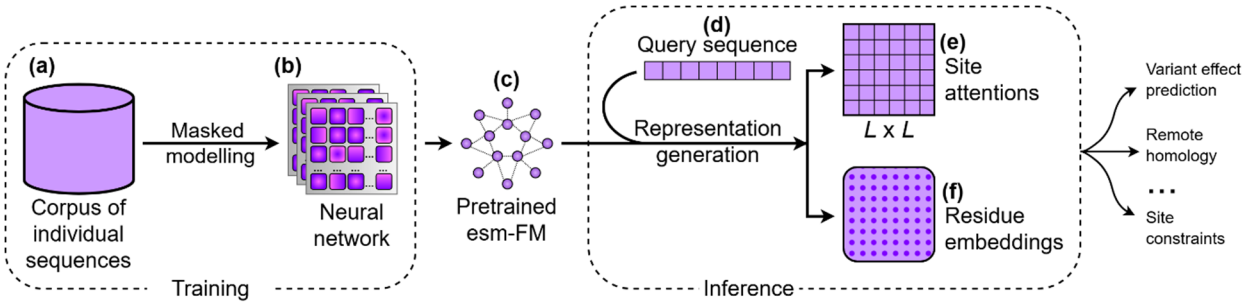


Figure 5. Training and using ESM foundation models

(**a**) ESM begins with individual sequences from an extensive database, which are used to train a transformer-based neural network using the masked modeling objective (**Box 1**). (**b**) The network shown comprises 33 layers, each with 20 attention heads, represented as squares. The pretrained model (**c**) can be used to generate a representation (**Box 1**) for any input sequence (**d**), yielding site attention matrices (**e**) and residue embeddings (**f**). These outputs can be used directly to make evolutionary inferences, such as residue contact maps[10,52], variant-effect predictions[53], and evolutionary relationships using sequence embeddings.[44]

Residue embeddings. In addition to attention scores, numerical representations (residue embeddings) that encode a residue's relationships to all other residues in the sequence can also be extracted from the pretrained esm-pFM (**Box 1**). These numerical representations of each residue function within the model integrate contextual evidence that informs predictions of residue identity. The final-layer embeddings are transformed to yield predicted probabilities for each residue. Masked positions with similar embeddings tend to result in similar predictions. Embeddings extracted from pretrained pFMs are biochemically meaningful: residues with similar physicochemical properties cluster together in embedding-based analysis.[10] For instance, acidic residues (D, E), aromatic residues (F, W, Y), and hydrophobic residues (L, I, M) form coherent groups despite the absence of any explicit chemical or evolutionary labels during training. Therefore, esm-pFMs can internalize core chemical principles purely from statistical regularities in the raw sequence data, without relying on predefined substitution matrices or alignments.

Sequence embeddings. The residue embeddings can be averaged element-wise to yield a sequence embedding, which is a fixed-dimensional representation (e.g., 1280 in ESM-1b[54]). Pairwise distances between sequence embeddings have been used directly to cluster sequences into families and to infer relationships[10,55], indicating that evolutionary relationships are preserved in this learned representation space. Therefore, foundation models can enable an alignment-free approach to construct protein or organismal phylogenies, a capability that could greatly expand large-scale evolutionary analyses. However, this potential remains to be tested for use in molecular systematics, and substantial work is needed to understand how variable sequence lengths, incomplete domain coverage, and embedding dimensionality affect such alignment-free inferences. Ultimately, while curated multiple sequence alignments will remain the gold standard for most evolutionary analyses, embedding-based approaches may offer valuable alternatives.

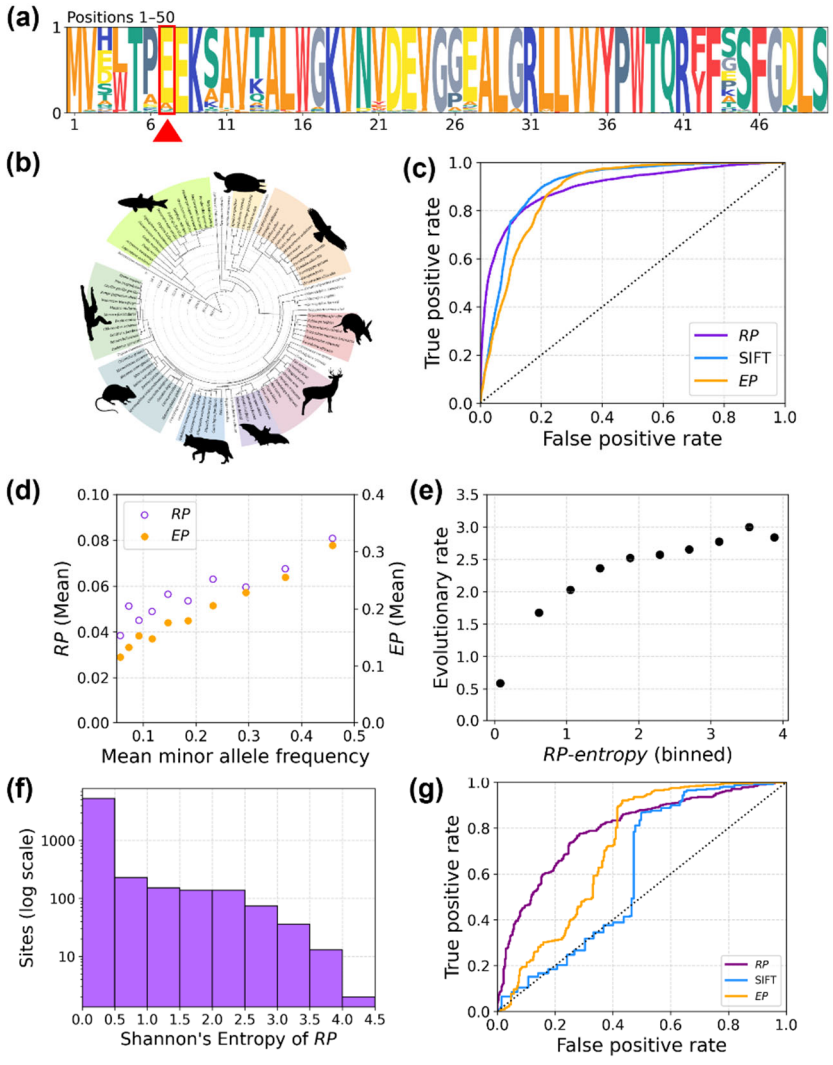
E. Imprints of the neutral theory of molecular evolution

A striking finding from the use of esm-pFM is that its predictions reflect the evolutionary patterns long associated with the neutral theory of molecular evolution.[24] Neutral theory posits that most of the observed variation among species is shaped by drift and purifying selection acting on functionally acceptable amino acids, and that population allele frequencies correlate with functional tolerability.[24,56] These principles were historically inferred from comparative genomics and population data, yet pFMs can recover them using only sequence statistics derived from individual sequences. The following demonstrates this convergence and shows how RepGen-derived residue predictions naturally encode neutrality, constraint, and tolerated variation.

RepGen analysis of a given protein sequence and pFM predicts a categorical probability distribution over the possible residues at each site[10,53], which we refer to as residue probabilities (*RPs*). *RP*s provide a direct, context-dependent measure of a residue's compatibility with the surrounding sequence. A high *RP* for a residue type indicates that the model identifies that type to be strongly favored by the sequence context. In contrast, low *RP* values signal biochemical or evolutionary implausibility. These probabilities are derived from patterns learned from millions of natural proteins, without requiring alignments or phylogenetic trees.

For example, in RepGen analysis using esm-pFM, the seventh residue of human HBB is predicted with high probability to be glutamic acid (E, *RP* = 0.88), whereas the sickle-cell disease-causing mutation to valine (V) receives an extremely low probability (*RP* = 0.01; **Fig. 6a**). This pattern generalizes across the proteome. Pathogenic missense variants cataloged in ClinVar[57] have much lower *RP*s (0.0032 ± 0.0002) than common human polymorphisms from gnomAD[58] (0.0564 ± 0.0012) in RepGen analysis (*see also*, ref.[53]). Consequently, even the direct use of *RP*s can achieve excellent performance in variant effect prediction (**Fig. 6c**).[53] The area under the Receiver operating characteristic curves (AUROC) for raw *RP*s derived from esm-pFM, which is comparable to established exclusively alignment-based approaches such as SIFT[30] (0.91) (**Fig. 6c**).

Figure 6. An analysis of human variants and polymorphisms using a single sequence pFM. (**a**) Residue probabilities (*RP*s) obtained using esm-pFM are shown in a logos representation for a portion of the human HBB sequence. The box highlights site 7, whose mutation (E→V) causes sickle cell anemia. (**b**) 100-species phylogeny obtained from the UCSC 100way Multiz Conservation Track, which was scaled to time using the TimeTree database[59] used for maximum likelihood analyses. (**c**) Receiver operating characteristic (ROC) curves for distinguishing pathogenic and neutral variants based on *RP*, SIFT, and *EP*. The area under the ROC curve (AUROC) for *RP*, SIFT, and *EP* is 0.90, 0.91, and 0.88. (**d**) Relationship between the population frequency of human polymorphisms and their *RP*s (closed circles) and *EP*s (open circles). Averages are plotted for 10 bins containing the same number of variants. (**e**) Relationship between *RP*-entropy derived from the distribution of *RP*s and traditional site-specific molecular evolutionary rates estimated using sequence alignments for sites carrying neutral variants, with each point representing the average for 10 bins with the same number of sites. (**f**) Distribution of *RP*-entropy for sites harboring common polymorphisms that have evolved with an evolutionary rate of 0. (**g**) The ROC curves for distinguishing pathogenic (6804) and neutral variants (401) at the slowest evolving sites. A balanced class size (10 replicates) was enforced when estimating ROC curves. Average AUROC is 0.80, 0.61, and 0.72 for *RP*, SIFT, and *EP*. See *Supplementary Methods* for calculation details.



This ability to distinguish pathogenic from neutral variants is not attributable to the inclusion of human protein sequence variants in the training set used to build esm-pFM. The training dataset, containing millions of protein sequences, was derived from the UniRef50 database, filtered for sequence quality and diversity. UniRef50 comprises a subset of UniProtKB protein sequences formed by clustering at 50% pairwise identity and retaining only a single representative sequence from each cluster. Such filtering eliminates closely related sequences and substantially reduces redundancy, thereby making human protein sequences constitute a relatively small fraction of the training data. Thus, the observed relationship between *RP*s and the neutrality of variants reflects general constraints predicted from patterns observed across diverse proteins, rather than from memorizing specific human variants. That is, the esm-pFM, trained solely on raw protein sequences without alignments or functional supervision, can recover classical pathogenicity signals, underscoring the neutral evolutionary information embedded in sequence statistics.[51,53]

The neutral theory also predicts that alleles at high population frequency tend to be more compatible with protein function and, thus, subject to weaker purifying selection.[56,60] This can be explored by comparing *RP* values for common amino acid polymorphisms stratified by allele frequency, which show that higher-frequency variants have higher *RP* values on average (**Fig. 6d,** closed circles) (*see also*, ref.[53,61]). The relationship between allele frequency and *RP* is similar to that seen for estimates of classical (neutral) evolutionary probabilities[56] (*EP*s, **Fig. 6d**), which are estimated by a Bayesian approach using multispecies alignments and a species tree (**Fig. 6b**). Yet, unlike *EP*s that depend on many explicit evolutionary assumptions enumerated earlier, *RP*s arise directly from the use of a pretrained esm-pFM. This concordance between *EP*s and *RP*s suggests that core signatures of drift and purifying selection, the primary mechanistic forces in the neutral theory, are implicitly encoded during masked modeling of sequences.

Sequence constraints versus historical conservation. The site-wise distribution of *RP*s also carries an imprint of neutral theory. For each site, we compute the Shannon entropy of *RPs* across all 20 amino acids predicted by esm-pFM[54], which we term *RP-entropy*. Low *RP-entropy* indicates a site where the model strongly prefers one or a few residues, whereas high *RP-entropy* marks sites where many residues are considered compatible.[54] Across large sets of proteins, *RP-entropy* correlates with absolute evolutionary rate estimates from traditional phylogenetic models (**Fig. 6e**) and ConSurf conservation scores[54], such that lower *RP-entropy* is associated with slower evolutionary rates at the site. This pattern supports a straightforward interpretation: esm-pFMs encode, in their probability distributions, the same constraint structure that classical rate-variation analyses infer from comparative sequence analyses. Crucially, this relationship emerges despite the absence of explicit phylogenies, branch lengths, or rate-category modeling in pretraining esm-pFM or in its use for RepGen inference.

However, this broad agreement in average values across bins hides substantial differences between site-specific *RP-entropy* values and evolutionary rates. For example, the distribution of *RP-entropy* is wide at the slowest evolving sites, with hundreds of sites showing large *RP-entropy* values (**Fig. 6f**). That is, multiple amino acid types are permissible at those sites. In contrast, the same amino acid type is found across species. Such discordance can arise from data sparsity, particularly when orthologous sequence alignments capture only a slice of evolutionary history, causing certain sites to appear slow-evolving, even though multiple amino acids are biochemically permissible. An additional possibility is that esm-pFMs, trained on a vast and diverse sequence

collection, can recognize broader biochemical flexibility even when it is not reflected in the assembled MSA.

In such cases, *RP-entropy* reflects possibility in the current context, rather than the historical outcomes captured by classical evolutionary rates. This could explain why 401 common polymorphisms (frequency > 5%) are found at the slowest evolving sites. For example, position 214 in the KIF25 protein, a member of the Kinesin family, shows no evolutionary substitutions over a 500-million-year history, but has a W214L polymorphism with a substantial allele frequency (17.8%). It has the highest *RP-entropy* (4.1) among the sites harboring these 401 polymorphisms. This observation suggests that reduced biological constraints in the current sequence context, despite a high degree of purifying selection at this site in the past and in other homologs within the family and/or other lineages, have permitted polymorphisms. Due to this property, the AUROC of *RP*s for diagnosing pathogenic versus neutral alleles at the highly conserved sites is good (0.80; **Fig. 6g**), as compared to SIFT and *EP* approaches that did not perform as well (0.61 and 0.72, respectively), consistent with previous findings that alignment based approaches tend to have a high false positive rates of variant effect diagnosis at highly conserved sites.[62]

This finding suggests that esm-pFMs may account for enigmatic patterns revealed by classical methods, because RPs depend on the entire sequence background rather than on residue frequencies across species. These patterns demonstrate that *RP*s provide a unique contemporary context-aware view of functional constraint, epistasis, and natural selection, which complements traditional comparative analyses that capture retrospective natural selection. Notably, however, not all differences between results from classical methods and deep learning are real. Estimation of millions to billions of parameters in deep learning can lead to artifacts due to the need to optimize performance during pretraining and may be the result of "over-fitting" or training-set biases[63,64], such as under-represented organisms. Therefore, unusual deep learning predictions should not be uncritically accepted or summarily dismissed.

Overall, however, these analyses suggest that the *RP* landscapes learned by esm-pFMs for sites and proteins are consistent with the neutral theory expectation that strongly deleterious variants are rare and concentrated at highly constrained sites, whereas common variants will occur primarily at more tolerant positions. Complementary work on unified biological foundation models that jointly encode DNA, RNA, and proteins shows that such FMs can recover central-dogma relationships between coding sequences and proteins and detect inconsistencies without explicit (or limited) supervised learning.[25] These converging lines of evidence support the view that FMs can rediscover core biological principles from raw sequence statistics.

F. Prospects for Evolutionary Inferences from Individual Sequences

Currently, molecular evolution and phylogenetic analysis depend on sequence alignments because comparing homologous sequences reveals substitutions that are necessary for estimating parameters of evolutionary models. With the introduction of pFMs, this can change, as the RepGen procedure produces residue embeddings, site attentions, and residue probabilities for a single sequence, enabling per-site evolutionary inference without explicit alignments for any protein. These estimates can be viewed as coming from a model of intramolecular epistasis that applies universally across protein families, analogous to Potts Hamiltonian models for individual families.[48,65] However, unlike Potts models, which are inferred from alignments of thousands of

homologous sequences, protein FMs are trained on heterogeneous, unaligned sequences from across the protein universe. This global training enables pFMs to learn from evolutionary constraints beyond a single protein family and to detect context-based dependencies without explicit alignments.[10]

That functional and structural constraints can be inferred from a single protein sequence using pFMs with billions of parameters, trained on millions of sequences[19,66] does not appear surprising in retrospect. Every new polypeptide in a cell is synthesized, folded, trafficked, and degraded by the same molecular machinery, governed by common biophysical rules for solubility, folding, binding, and regulation. These rules are implemented mechanistically by the cellular environment, but they leave a statistical fingerprint on sequences: residue types and combinations that systematically violate them are purged by purifying selection, whereas compatible ones recur across individuals, organisms, proteins, and families.

Consequently, pFMs can provide useful evolutionary and functional readouts for any protein, including proteins with shallow or no detectable homology, such as remote or 'orphan' proteins for which good alignments or phylogenies are challenging to obtain.[67–70] This is important because many proteins of interest may be lineage-specific, highly diverged, or composed of rare domain architectures that defy confident homology detection. In such cases, traditional methods may struggle to estimate substitution rates, detect selection, or assess the impact of mutations, because they depend on multiple alignments of homologous sequences. Additionally, pFMs extend the reach of molecular evolutionary analysis by enabling analysis beyond missense variation, as the same RepGen analysis can, in principle, be applied to other classes of protein variants, including stop-gain and insertion-deletion variants[53,71] that have been difficult to model in standard evolutionary approaches.

In general, however, it will be prudent to investigate pFM-derived results and hypotheses using alignment-based conservation measures, phylogenetic reconstructions, and functional or structural assays. Discrepancies between pFM-based and traditional inferences will be especially informative, as they may highlight cases where pFM has captured constraints that are not apparent from current sequence samples, or conversely, where data gaps or training biases limit pFM's reliability. Treating pFMs as augmenting classical workflows will allow researchers to exploit their speed and breadth while maintaining the rigor of comparative, alignment-based analysis. Extensive and systematic benchmarking is needed to develop guidelines for when to trust pFM-derived inferences and when they outperform traditional approaches.

G. Feasibility of using foundation models in everyday analyses

An important practical question is the feasibility and computational cost of applying deep learning in routine data analysis. Here, it is essential to distinguish between the substantial one-time cost of training a pFM and the relatively modest RepGen procedure required to make inferences with the pretrained model. Training state-of-the-art pFMs typically requires many GPU-days on specialized hardware clusters and incurs nontrivial financial and energy costs; however, this expense is borne once by model developers rather than by downstream users. Once a pretrained model is made publicly available, it can be used for RepGen analysis, whose computational cost is the primary concern in standard molecular evolutionary investigations.

To examine this, we measured the time required to run RepGen with an esm-pFM (≈650M parameters) on a collection of short and long protein transcripts. RepGen runtimes were on the order of seconds, even on commodity GPUs (e.g., RTX 3050) found in laptops costing a few hundred dollars (**Fig. 7**). The relationship between protein length and runtime is close to linear, indicating that esm-pFM analysis with a model containing hundreds of millions of parameters is tractable on commodity GPUs (as well as multi-core CPUs). In contrast, models with billions of parameters require substantially more video memory, because GPU memory usage scales approximately linearly with the number of model parameters and roughly quadratically with the processed sequence length. Fortunately, many optimizations and heuristics are being developed to enable the use of larger and more accurate models even on commodity hardware.[72–74] With these advances, we anticipate that pFM-based workflows will soon allow researchers to obtain rapid, first-pass assessments of proteins, domains, and variants before committing to more labor-intensive alignment-based pipelines (**Fig. 1**).

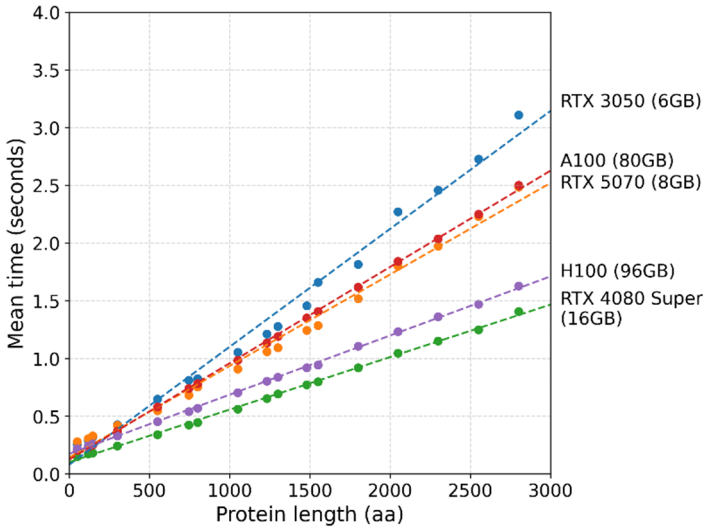


Figure 7 | Runtime of RepGen analysis using an esm-pFM with 650 million parameters.
RepGen calculations were performed on 13 protein sequences, ranging from 50 to 2850 residues in length, using five different GPU types. Each point represents the mean of 10 RepGen runs for a given protein, and the best-fit scaling curve for time-to-completion was approximately linear with respect to sequence length (R^2 = 0.99 for all five curves). In December 2025, representative US purchase prices were roughly $200 for an RTX 3050, $550 for an RTX 5070, $1,800 for an RTX 4080, $15,000 for an A100, and $25,000 for an H100, illustrating that useful RepGen throughput is attainable even on commodity GPUs, with further speedups on data-center accelerators. The VRAM size is shown for each GPU.

H. Conclusions and outlook

Protein foundation models represent a conceptual and practical shift in how and what is studied in molecular evolution. Classical substitution models, built from counts of observed changes in MSAs, offer interpretable summaries of past substitution tendencies and rate variation, but they treat evolutionary patterns across sites and lineages as independent and identically distributed. In contrast, pFMs are large, single models trained on vast collections of sequences, whose parameters encode high-dimensional patterns of residue compatibility and context-dependent constraints across the entire observed protein universe. These pFMs are far from black boxes. Their internal components and outputs are biologically interpretable and evolutionarily meaningful. Sequence attention maps reveal phylogenetic relationships; site attention maps show structural and functional couplings; residue embeddings identify pathogenic variants and recover site-specific constraints; and sequence embeddings show evolutionary relatedness across

sequences. These signals arise not from explicit evolutionary models, but from the ability to identify patterns in the data during self-supervised training.

A striking insight is the correspondence between pFM behavior and the neutral theory of molecular evolution. Residue probabilities produce results consistent with classical expectations relating allele frequency, constraint, and mutational tolerance, despite the absence of phylogenetic trees or evolutionary assumptions. That such principles emerge from raw sequences alone suggests that the statistical patterns of protein space already encode the imprint of drift and purifying selection, and that pFMs provide a massive data-driven realization of these evolutionary forces. This ability is further demonstrated by generalized FMs trained on both nucleotide and protein sequences that encode the central dogma of molecular biology: DNA and protein embeddings from the same gene converge in embedding space, even though they were not paired during training.[25]

Furthermore, FMs do not simply approximate classical approaches and can produce significantly different results for the same protein or residue. Undoubtedly, molecular evolution is not merely the accumulation of independent substitutions, but the evolution of contextual residue interactions that permit substitutions at individual residues over the millennia. Classical methods, with their simplifying assumptions, seem to model historical patterns well, which is sufficient in many situations. Instead, FMs focus on residue dependencies that are essential for proper protein function within an organism; therefore, evolutionary analysis of embeddings and attention maps produced by FMs can help reveal the dynamic evolution of residue relationships in the past as well as contemporary constraints. With this capability, pFMs are poised to offer more mechanistic explanations that go beyond simply describing observed residue differences arising from the evolution of these inter-residue dependencies under purifying selection.

Although pFMs offer powerful new capabilities, they also have significant limitations. They do not encode explicit evolutionary timescales, substitution histories, or branch-length information. Their performance depends on the statistical composition of their training datasets, which may underrepresent certain protein types or rapidly evolving lineages.[75] *RP*s and attention maps reflect the model's estimate of contextual plausibility, not the realized evolutionary path taken by any lineage. Furthermore, pFMs can overgeneralize when analyzing sequences with domain-specific architectures or residue environments that are underrepresented in the training protein collection. Recognizing and investigating these limitations is essential when interpreting pFM-based evolutionary inferences.

By design, pFMs are generalists, but no single model can excel at all downstream tasks. Just as Dayhoff-style substitution matrices were later specialized for mitochondrial, function-, or organism-specific proteins, tailoring pFM training data and fine-tuning them can yield superior results for particular applications.[76,77] For instance, ESM-2[50] was optimized for protein structure prediction, and ESM-1v[61] for variant-effect prediction, each benefiting from domain-specific sequence collections and other optimizations. Such sensible pFMs mitigate the risk of model homogenization, in which mismatched training data can propagate systematic errors.[78]

As pFMs continue to expand in scale and generality, and as interpretability tools improve, they are becoming increasingly valuable to evolutionary biology[79,80] and phylomedicine[6]. This development is timely: while protein sequence databases continue to grow rapidly, our

understanding of the functions and origins of many proteins remains limited. Advances in pFMs offer a promising approach to closing this gap, thereby enhancing our understanding of the growing universe of protein sequences. In these efforts, classical models provide interpretability, whereas pFMs offer mechanistic depth. Together, they can create a unified framework for understanding protein evolution that integrates biological principles with data-driven discovery. By combining mechanistic insights with large-scale data modeling, the field can uncover not only how proteins have evolved but also how evolutionary constraints and dependencies influence the range of viable sequences today. Protein foundation models thus present a new perspective, grounded in data yet aligned with evolutionary theory and molecular biology, for understanding the forces that shape molecular evolution.

Acknowledgements

We thank Sudip Sharma, Qiqing Tao, Hardik Sharma, and S. Blair Hedges for their many critical and helpful comments on an earlier draft of this manuscript. We thank Jack Craig for assistance in calibrating the UCSC Multiz phylogeny. This work was supported in part by a grant from the National Institutes of Health (R35GM139540-05). Part of the calculations were carried out on HPC resources supported in part by the National Science Foundation (MRI 1625061) and the US Army Research Laboratory (W911NF-16-2-0189).

Bibliography

1. EMBL-EBI. Current Release Statistics. https://www.ebi.ac.uk/uniprot/TrEMBLstats.
2. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
3. UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
4. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
5. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 1–8 (2018).
6. Kumar, S., Dudley, J. T., Filipski, A. & Liu, L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet.* **27**, 377–386 (2011).
7. Kumar, S. Embracing green computing in molecular phylogenetics. **39**, msac043 (2022).
8. Pollock, D. D., Taylor, W. R. & Goldman, N. Coevolving protein residues: maximum likelihood identification and relationship to structure. *J. Mol. Biol.* **287**, 187–198 (1999).
9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2016239118 (2021).
11. de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).
12. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*. (Oxford University Press, New York, 2000).
13. Yang, Z. *Computational Molecular Evolution*. (Oxford University Press, London, England, 2006).
14. Rao, R. *et al.* MSA Transformer. *Proceedings of the 38th International Conference on*

Machine Learning **139**, 8844–8856 (2021).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for language understanding. *arXiv [cs.CL]* (2018).
16. Vig, J. *et al.* BERTology meets biology: Interpreting attention in protein language models. *arXiv [cs.CL]* (2020) doi:10.48550/ARXIV.2006.15222.
17. Lupo, U., Sgarbossa, D. & Bitbol, A.-F. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nat. Commun.* **13**, 6298 (2022).
18. Moreno-Muñoz, P., Recasens, P. G. & Hauberg, S. On masked pre-training and the marginal likelihood. *arXiv [stat.ML]* (2023).
19. Hou, C., Liu, D., Zafar, A. & Shen, Y. Understanding language model scaling on protein fitness prediction. *bioRxivorg* (2025) doi:10.1101/2025.04.25.650688.
20. Levy, R. M., Haldane, A. & Flynn, W. F. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* **43**, 55–62 (2017).
21. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. in *Atlas of Protein Sequence and Structure (ed. Dayhoff, M. O.)* 345–352 (National Biomedical Research Foundation, 1978).
22. Vaswani, A. *et al.* Attention is all you need. *arXiv [cs.CL]* (2017).
23. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
24. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, Cambridge, England, 1985).
25. He, Y. *et al.* Generalized biological foundation model with unified nucleic acid and protein language. *Nat. Mach. Intell.* **7**, 942–953 (2025).
26. Brixi, G. *et al.* Genome modeling and design across all domains of life with Evo 2. *bioRxiv* (2025) doi:10.1101/2025.02.18.638918.
27. Zhang, X., Yang, M., Yin, X., Qian, Y. & Sun, F. DeepGene: An efficient foundation model for genomics based on Pan-genome graph transformer. *bioRxiv* (2024) doi:10.1101/2024.04.24.590879.
28. Dayhoff, M., Schwartz, R. & Orcutt, B. C. Matrices for detecting distant relationships. *Atlas of protein sequence and structure* (1978).
29. Altschul, S. F. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555–565 (1991).
30. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
31. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8**, 275–282 (1992).
32. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 10915–10919 (1992).
33. Tamura, K. & Kumar, S. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.* **19**, 1727–1736 (2002).
34. Moran, R., Morgan, C. & O'Connell, M. A guide to phylogenetic reconstruction using heterogeneous models—A case study from the root of the placental mammal tree. *Computation (Basel)* **3**, 177–196 (2015).
35. Baele, G. Context-dependent evolutionary models for non-coding sequences: An overview of several decades of research and an analysis of laurasiatheria and primate evolution. *Evol. Biol.* **39**, 61–82 (2012).
36. Bhattacharya, N. *et al.* Interpreting Potts and transformer protein models through the lens of simplified attention. in *Biocomputing 2022* (WORLD SCIENTIFIC, 2021). doi:10.1142/9789811250477_0004.
37. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **15**, 816–822 (2018).
38. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary

data. *Nature* **599**, 91–95 (2021).
39. Szałata, A. *et al.* Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).
40. Rao, R. *et al.* MSA Transformer. *Proceedings of Machine Learning Research* 8844–8856 (2021).
41. Nesterenko, L., Boussau, B. & Jacob, L. Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. *bioRxiv* (2022) doi:10.1101/2022.06.24.496975.
42. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
43. Chen, R., Foley, G. & Bodén, M. Learning the language of phylogeny with MSA Transformer. *Cell Syst.* 101445 (2025).
44. Tule, S., Foley, G. & Bodén, M. Do protein language models learn phylogeny? *Brief. Bioinform.* **26**, (2024).
45. Nesterenko, L., Blassel, L., Veber, P., Boussau, B. & Jacob, L. Phyloformer: Fast, accurate, and versatile phylogenetic reconstruction with deep neural networks. *Mol. Biol. Evol.* **42**, (2025).
46. Yax, N., Oudeyer, P.-Y. & Palminteri, S. PhyloLM : Inferring the phylogeny of Large Language Models and predicting their performances in benchmarks. *arXiv [cs.CL]* (2024).
47. Khatri, K., Levy, R. M. & Haldane, A. Phylogenetic corrections and higher-order sequence statistics in protein families: The Potts model vs MSA Transformer. *ArXiv* (2025).
48. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 67–72 (2009).
49. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **87**, 012707 (2013).
50. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
51. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
52. Zhang, Z. *et al.* Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2406285121 (2024).
53. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
54. Marquet, C. *et al.* Embeddings from protein language models predict conservation and variant effects. *Hum. Genet.* **141**, 1629–1647 (2022).
55. Dickson, A. & Mofrad, M. R. K. Fine-tuning protein embeddings for functional similarity evaluation. *Bioinformatics* **40**, (2024).
56. Liu, L., Tamura, K., Sanderford, M. D., Gray, V. E. & Kumar, S. A molecular evolutionary reference for the human variome. *Mol. Biol. Evol.* **33**, 245–254 (2016).
57. Landrum, M. J. *et al.* ClinVar: updates to support classifications of both germline and somatic variants. *Nucleic Acids Res.* **53**, D1313–D1321 (2025).
58. Chen, Siwei Francioli, Laurent C Goodrich, Julia K *et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
59. Kumar, S. *et al.* TimeTree 5: An expanded resource for species divergence times. *Mol. Biol. Evol.* **39**, (2022).
60. Kumar, S. & Patel, R. Neutral theory, disease mutations, and personal exomes. *Molecular biology and evolution* **35**, 1297–1303 (2018).
61. Meier, Joshua Rao, Roshan Verkuil, Robert Liu, Jason Sercu, Tom Rives, Alexander. Language models enable zero-shot prediction of the effects of mutations on protein

function. *35th Conference on Neural Information Processing Systems* (2021).
62. Kumar, S., Sanderford, M., Gray, V. E., Ye, J. & Liu, L. Evolutionary diagnosis method for variants in personal exomes. *Nat. Methods* **9**, 855–856 (2012).
63. Ye, W. *et al.* The clever Hans mirage: A comprehensive survey on spurious correlations in machine learning. *arXiv [cs.LG]* (2025).
64. Shah, M. & Sureja, N. A comprehensive review of bias in deep learning models: Methods, impacts, and future directions. *Arch. Comput. Methods Eng.* **32**, 255–267 (2025).
65. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1293–301 (2011).
66. Weissenow, K. & Rost, B. Are protein language models the new universal key? *Curr. Opin. Struct. Biol.* **91**, 102997 (2025).
67. Lamb, K. D. *et al.* From a single sequence to evolutionary trajectories: protein language models capture the evolutionary potential of SARS-CoV-2 protein sequences. *bioRxiv* (2024) doi:10.1101/2024.07.05.602129.
68. Singh, J., Litfin, T., Singh, J., Paliwal, K. & Zhou, Y. SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **38**, 1888–1894 (2022).
69. Jing, X., Wu, F., Luo, X. & Xu, J. Single-sequence protein structure prediction by integrating protein language models. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2308788121 (2024).
70. Kabir, A., Moldwin, A., Bromberg, Y. & Shehu, A. In the twilight zone of protein sequence homology: do protein language models learn protein structure? *Bioinform. Adv.* **4**, vbae119 (2024).
71. Hegde, M., Nebel, J.-C. & Rahman, F. Language modelling techniques for analysing the impact of human genetic variation. *Bioinform. Biol. Insights* **19**, 11779322251358314 (2025).
72. Sgarbossa, D., Malbranke, C. & Bitbol, A.-F. ProtMamba: a homology-aware but alignment-free protein state space model. *Bioinformatics* **41**, (2025).
73. Tay, Y., Dehghani, M., Bahri, D. & Metzler, D. Efficient Transformers: A survey. *ACM Comput. Surv.* **55**, 1–28 (2023).
74. Schmidinger, N. *et al.* Bio-xLSTM: Generative modeling, representation and in-context learning of biological and chemical sequences. *arXiv [q-bio.BM]* (2024) doi:10.48550/ARXIV.2411.04165.
75. Ding, F. & Steinhardt, J. Protein language models are biased by unequal sequence sampling across the tree of life. *bioRxiv* 2024.03.07.584001 (2024) doi:10.1101/2024.03.07.584001.
76. Sledzieski, S. *et al.* Democratizing protein language models with parameter-efficient fine-tuning. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2405840121 (2024).
77. Schmirler, R., Heinzinger, M. & Rost, B. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications* **15**, (2024).
78. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv [cs.LG]* (2021).
79. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **13**, 274–285.e6 (2022).
80. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3 (2021).
81. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
82. Kumar, S. *et al.* MEGA12: Molecular Evolutionary Genetic Analysis version 12 for adaptive and green computing. *Mol. Biol. Evol.* **41**, (2024).

Supplementary Methods.

Many studies have reported relationships among various measures derived from transformer models and traditional molecular evolutionary methods, as well as with the frequency of variants in humans.[51,53,56] They have used different, albeit overlapping, datasets and approaches. For uniformity and to enable direct comparison of patterns across panels in this perspective article, we have reproduced those trends using a common dataset and techniques, as follows.

Multiple Sequence Alignments (MSAs). The database was compiled using the UCSC 100way Multiz Conservation Track as the initial reference point.[81] The Multiz Track comprises nucleotide and amino acid sequence alignments for 100 vertebrate species that are putatively orthologous. We used the knownCanonical subset of the Multiz track, which restricts the alignment pool to alignments representing the target protein's canonical isoform. The knownCanonical subset comprises 23,335 multiple sequence alignments (MSAs) spanning nearly the entire genome, from chr1:169888676 to chrY:19077367. We used the corresponding protein sequences for each DNA alignment.

Pathogenic and neutral variants. We selected all missense variants reported in the ClinVar database[57] with the following annotations: “pathogenic”, “likely pathogenic”, and “pathogenic/likely pathogenic”. From this collection, we only retained alleles with a global minor allele frequency < 0.5% to build a clear-cut Pathogenic variant dataset. This yielded 13,359 pathogenic variants. To construct a neutral variant database under the assumptions of the neutral theory of molecular evolution, we used the gnomAD database[58] and selected all common missense variants with global frequency between 5% and 50%. We excluded all variants in which the gnomAD-specified alternative allele was not the minor allele. Applying these criteria identified 14,634 neutral variants.

Residue Probabilities (RPs) and Entropy. RepGen analysis of the human protein sequence used ESM-2[50] pFM. All *RP* calculations were carried out on an Nvidia A100 GPU using the entire amino acid sequence as the input, except for the titin protein (TTN). TTN (ENST00000360870) comprises 35,592 residues, necessitating a sliding-window (1500-residue) approach for RepGen analysis. Shannon’s entropy of the *RP* distribution was calculated using the following formula, as implemented in Python: $-\Sigma p_i \log_2(p_i)$, where p_i is the probability of each residue type and the sum is over 20 residues.

Sequence Attentions and Phylogeny Inference. Following a protocol[43] for msat-pFM phylogenetic inference, we inferred phylogenies by computing a pairwise attention distance matrix in which the Euclidean distances between a pair of sequences (*p* and *q*) were based on their attention maps, as follows:

$D_{pq}^{(l)} = \sqrt{\sum_{j=1}^{L} \sum_{k=1}^{d} (X_{pjk}^{(l)} - X_{qjk}^{(l)})^2}$, where l is the number of layers in the model, L is the number of column attention heads in the layer, and d is the number of residue matrices for each column.

We used the pairwise distance matrices from each attention head to calculate a mean distance matrix. We applied the neighbor-joining algorithm to this mean distance matrix to obtain a phylogeny for each input alignment.

Evolutionary Probabilities (EPs) and SIFT scores. *EP*s[56] were calculated in MEGA 12[82] using the 100-species timetree (**Fig. 6b**), uniform rates across sites, and a Poisson model of amino acid substitutions to avoid infusing biochemical and rate variation information in *EP* inference. SIFT scores[30] were imported from gnomAD.[58] SIFT and *EP* were selected because they represent distinct methods for estimating residue neutrality and rely exclusively on evolutionary information, unlike many other approaches that also incorporate population frequencies and/or biochemical and structural data.

Estimation of classical evolutionary rates. For each MSA, we used the Maximum Likelihood method to first estimate rates among sites under an LG+F substitution model using the species tree (**Fig. 6b**) in MEGA 12 software[82]. Relative site rates were converted to absolute rates using the total number of substitutions estimated for each MSA and the divergence times extracted from the TimeTree resource [59].

Receiver operating characteristic (ROC) curve analysis. For all analyses, we constructed balanced test sets to prevent artifacts from class imbalance. Specifically, for each subset of interest, we subsampled the majority class without replacement to match the minority class. The reported ROC curves and AUROCs are averages across 10 randomly balanced sets.

A table of variants and associated estimates is provided in the *Supplementary File*.