PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Descent from a common ancestor restricts exploration of protein sequence space

Lada H. Isakova[a], Elizaveta Streltsova[b], Olga O. Bochkareva[c], Peter K. Vlasov[d], and Fyodor A. Kondrashov[a,1]

Affiliations are included on p. 10.

Edited by Eugene V. Koonin, National Institutes of Health, Bethesda, MD; received November 16, 2025; accepted March 3, 2026

How functional protein sequences are distributed in sequence space is fundamentally important for evolutionary theory and protein design, particularly if a large diversity of protein functions are hidden in evolutionarily unexplored areas of the sequence space. However, this question is understudied in part because experimental and computational studies use extant sequences as a starting point to study sequence space. Here, we study whether extant sequences are representative of the entire functional sequence space. Across thousands of protein families from vertebrates and bacteria we calculate the dimensionality and the volume of sequence space occupied by extant homologs. We find that the observed dimensionality and volume of extant sequence space are minuscule, many orders of magnitude smaller than what we estimated using a model of protein evolution. Simulating sequence evolution we then quantify the impact of phylogeny, selection, and epistasis on restricting the evolutionary exploration of sequence space. We find that sequence evolution from a single common ancestor, or a single point of origin in sequence space, is by far the largest limiting factor that reduces the dimensionality and volume of extant sequence space. These results indicate that there are vast areas of functional sequence space that have not been explored in evolution because of the excessive restrictions on natural exploration of the protein sequence space imposed by the point of origin effect. We suggest that protein design methods that rely on extant sequences may be limited in their ability to discover truly novel functions.

protein evolution | sequence space | protein function | dimensionality

Significance

Are natural protein sequences representative of all possible sequences that are functional? The sequence space is immense but proteins have been evolving for billions of years, so much of the possible functional space may have already been explored. We find that because sequence evolution of homologous proteins starts from a single common ancestor, protein sequence diversity has been limited to an extreme degree and even 4 billion years of evolution was insufficient to explore the functional sequence space. Protein engineering models that learn only from natural protein sequences may be limited in their ability to predict sequences outside the explored sequence space and empirical approaches that explore the unnatural sequence space may be necessary to fully realize their potential.

The advent of complex statistical data analysis and AI is changing how we study and design protein molecules (1). Among the successes of the application of AI is its ability to accurately predict protein structure from sequence (2, 3). However, AI tools have not yet reached the same level of success when tasked with predicting higher-level functions, such as enzymatic activity or fluorescence (1). Existing approaches can accurately predict the functionality of sequences that contain several mutations relative to known natural sequences (4). However, despite individual success stories (5, 6), even function-specific models cannot accurately predict functional sequences with many different mutations relative to known sequences (7). A common approach is to train AI models on all natural protein sequences (8) and have them predict sequences with a desired property that have not occurred in evolution, in a similar manner that AlphaFold predicts structure. It is unclear to what extent this approach may work because even the vast numbers of available natural sequences represent a small proportion of all possible sequences (1).

AI models that were trained on natural sequences may be able to generalize more broadly if natural sequences are representative of the properties of functional sequences as a whole (9). Natural sequences are the products of molecular evolution, thus, an evolutionary context of how natural sequences evolve is important for understanding the relevance of natural sequences for generalized AI model training. Depending on the nature of the fitness landscape, the distribution of natural sequences may be either rather limited, or be vast and contain enough information for AI models to generalize across the entire sequence space. In evolutionary biology, these issues have been studied for almost a century. Thus, we provide a brief review of the issue of natural sequence distribution from the perspective of evolutionary theory.

Sequence space—an abstract space of all possible amino acid sequences—is high-dimensional and astronomically large (10, 11). Natural, extant protein sequences of any given protein family are somehow distributed in this sequence space (12, 13), and this distribution is the result of evolution from ancestral sequences (14). Over the course of evolution protein sequences must remain functional to contribute to organisms' fitness. Thus, the distribution of extant sequences in sequence space forms a continuous network of high fitness (15–19) that corresponds to the phylogenetic trajectory of how the protein

Author contributions: L.H.I., E.S., O.O.B., P.K.V., and F.A.K. designed research; L.H.I., E.S., O.O.B., P.K.V., and F.A.K. performed research; L.H.I., E.S., O.O.B., and F.A.K. analyzed data; and L.H.I. and F.A.K. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: fyodor.kondrashov@oist.jp.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2532018123/-/DCSupplemental.

Published March 31, 2026.

sequences evolved in that sequence space. While the mechanisms governing how sequences evolve have been the subject to extensive scrutiny (20, 21), less attention has been given to how molecular evolution shapes the resulting distribution of extant sequences in sequence space (refs. 13, 14, and 22–25).

Due to the multidimensional and abstract nature of the sequence space, it is prudent to define terms and construct a mental picture of the subject of our research. The entire sequence space is an abstract representation of all possible protein sequences. However, protein length can vary extensively, so it is not convenient to project the smallest and the largest proteins into the same sequence space. Therefore, here we consider separately the sequence space of individual protein families consisting of alignable homologues. Thus, for an alignment of homologous sequences with a roughly constant length L, the protein family can be represented in an L-dimensional space with 20^L possible sequences.

Not all of these 20^L sequences are functional, indeed, a vast majority of them are not. Conversely, functional sequences are extremely rare, with estimates of the proportion of functional sequences for a handful of protein families range from 10^{-12} to 10^{-126} (26). From an evolutionary perspective, functional protein sequences can be classified into four different types (Fig. 1*A*): 1) extant sequences, those found in currently living organisms (green dots), some of which may have already been sequenced; 2) ancestral sequences (orange phylogeny), which were evolutionary precursors to extant ones; 3) extinct sequences, those which have not led to any extant descendents (red branches on phylogeny); and 4) functional sequences that never existed (blue area).

When the fitnesses of all sequences are known, the resulting shape is often referred to as the fitness landscape (10, 11, 18) because it can be characterized by low-fitness valleys and high-fitness ridges. Understanding the shape of the entire fitness landscape, therefore, requires a definition of fitness for all sequences in sequence space, regardless of whether or not those sequences are functional or have ever existed. Complex fitness ridges emerge among the fitness valleys due to epistasis, which, at the most basic level, is the interaction between different allele (amino acid) states at different loci (sites). Mathematically speaking, a nonepistatic fitness function is e^P, where $P = x_1 + x_2 + \ldots + x_n$, x_n is the fitness contribution of allele x_n at site n. The reason why an exponential function represents a nonepistatic case is because it ensures that each allele impacts fitness independently from all other alleles at other loci, so that the effect of an allele that increases fitness by 10% in one genetic context will have exactly the same effect in all other contexts.

Conversely, any deviation from this exponential function is defined as epistasis, with simple deviations, such as faster or slower than exponential growth representing simple, or unidimensional, epistasis. When a unidimensional function cannot be used to define the fitness of any genotype in sequence space the fitness landscape has fitness valleys and ridges (11, 18, 19, 28–31) driven by complex, multidimensional forms of interaction between alleles at different sites. Maynard Smith's powerful analogy (15) for sequence space helps to build an intuition of how context dependence leads to complex fitness landscapes (32, 33). Several trajectories between English words WORD and GENE can be made, whereby only a single letter is changed at a time and only existing words in the English language are acceptable, i.e., have high fitness. These trajectories represent ridges of high fitness, surrounded by nonsensical four letter word combinations, representing the fitness valleys. The contribution of different letters at each position is highly context-dependent; for example, the substitution of O –> E at the second site has low fitness in the context of W _ RD but has high fitness in the context of G _ NE. Such context dependence, frequently detected in protein sequences (34), shapes the fitness landscape in interesting and nontrivial ways (35). The fitness landscape may contain independent, unconnected fitness ridges (36, 37) surrounded by vast valleys of low fitness. The fitness ridges themselves may be large but porous, or "holey" (38). Thus, the functional sequence space may be sparse, i.e., not monolithic, such that most functional sequences have more nonfunctional sequence neighbors than functional ones, and yet functional sequences may be found at substantial distances from each other in sequence space (18, 28, 34).

Three empirical approaches can be used to ascertain the nature of the fitness landscape. The first is to experimentally measure the fitness or function of a set of sequences with deep mutational scanning (39, 42, 43) or directed evolution [(44, 45) and references within]. The second is to use extant sequences to calculate a set of parameters to infer the nature of the fitness landscape (34, 46, 47). Finally, one can use extant sequences to build generative computational models of the fitness landscape. Explicit models such as Potts or Direct Coupling Analysis (48–50) specify fitness as an interpretable function of site-specific amino acid frequencies

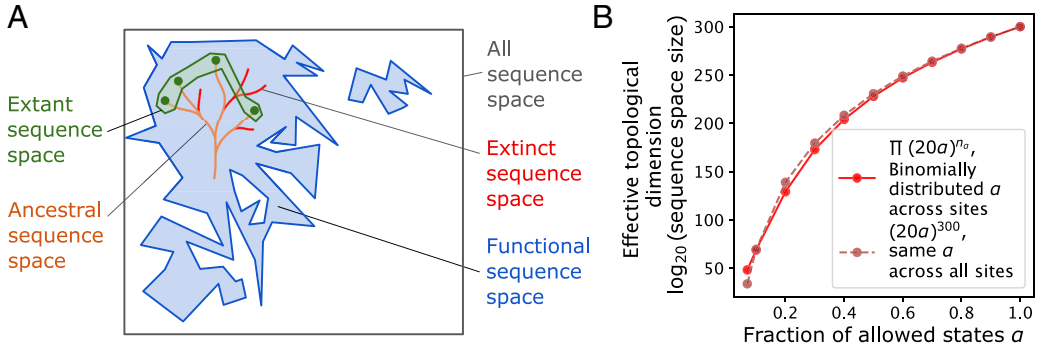

Fig. 1. The dimensionality and volume of sequence space. (*A*) Conceptual representation of the four different types of functional sequences. In sequence space (black square), extant sequences (green) represent a tiny subset of functional sequences that occurred in evolution (orange and red phylogeny), and even smaller than the unknown subset of all possible functional sequences (blue area). The green outline represents the space defined by extant sequences. (*B*) Maximum expected dimensionality of functional sequence space as a function of the fraction of allowed amino acid states, α. The effective topological dimension is calculated as the $\log_{20}$ of possible number of sequences of length 300 with 20α possible amino acids per site (brown dashed line) and a binomially distributed amino acid usage across sites with mean 20α (red solid line). The dashed line is adapted from (27) with x axis changed from number of amino acid types to α.

and pairwise couplings, while implicit ones, including variational autoencoders (51) and the variety of protein language models (1, 52) learn high-dimensional representations of sequence space, from which fitness can be inferred.

In practice, all of these approaches characterize local areas of the fitness landscape much better than they can extrapolate to larger, global fitness landscapes. The definitions of local and global areas of the fitness landscape are not precise, nevertheless, these terms are useful in a variety of contexts (53). A local area of the fitness landscape signifies a specific limited area of the sequence space. For example, experimental assays that survey the function of genotypes in proximal vicinity to a specific sequence survey the local fitness landscape (ref. 41). A truly global fitness landscape would encompass the entire sequence space of all possible proteins. The complete sequence space of an entire protein, or protein family, is also a rather global fitness landscape. However, no clear threshold of similarity that would distinguish a local and a global landscape exists, and a formalistic definition, such that a local landscape is an area within 10% sequence divergence from some point, is not useful. Yet we understand that the shape and properties of the local and global fitness landscapes may be quite different (7, 40, 54). It is also conceivable that some intermediate fitness landscape scale between the local and global may have its own utility. Perhaps a useful way to think about fitness landscapes is to consider a continuum of locality, starting from immediate single-nucleotide neighbors as the most local space that expands to the entire sequence space on the ultimate, global end of this continuum (55, 56).

Deep learning and AI-driven approaches have the best track record in predicting functional sequences from available data (5, 6, 40, 42, 57–60), in part because they are designed to depart from the study of the local areas of the fitness landscape restricted to extant sequences and a few of its experimental derivatives (7). However, these approaches are in their infancy and have not yet led to robust descriptions of global features of the fitness landscape. Moreover, despite certain ability to generalize, even most sophisticated generative models are biased toward their training data which are limited to the evolutionary sequence records (1, 61). To what extent the reliance of training on natural sequences fundamentally limits these models depends, on the one hand, on the interplay of selection and epistasis, and, on the other hand, on evolutionary contingency of nature having explored a sufficient area of the sequence space. If extant sequences occupy a large area that is representative of the sequence space as a whole, it may be possible to create generalizable models from natural sequence data. By contrast, if protein sequence space is defined by complex fitness landscapes and natural sequences do not reveal all possible aspects of complex epistatic interactions of that landscape then models relying on natural sequences may never be able to generalize using natural sequence data alone.

Thus, there is a persistent need to elucidate the global properties of the fitness landscape. As a case in point we have hardly progressed in answering the four fundamental questions about the fitness landscape that were formulated by John Maynard Smith half a century ago (15): 1) Are all extant sequences located on the same fitness ridge? 2) Did evolution pass through fitness valleys? 3) What fraction of functional sequences has been explored by evolution? 4) What fraction of functional sequences are inaccessible? Here, we consider the geometric properties of the sequence space occupied by extant sequences to determine whether natural sequences are sufficiently representative of the global fitness landscape.

Results

Estimating Dimensionality of Extant Sequence Space. We follow a similar approach employed previously on a smaller number of sequences (23, 55) and calculate the intrinsic dimension of extant sequence distribution in sequence space. In discrete sequence space every protein site corresponds to a dimension axis with 20 possible states (amino acids) with the maximum dimensionality equal to the total sequence length. The effective dimensionality of the occupied sequence subspace can be obtained by calculating the intrinsic dimension—a measure of the number of axes that describe the underlying space. Correlation dimension (D_{cor}) is one of the estimators of the intrinsic dimension and is calculated by measuring how the number of pairs of points $N(r)$ within a certain distance r grows as that distance increases: $N(r) \sim r^{D_{cor}}$ (62, 63) (Fig. 2 and *SI Appendix, Methods*). As a distance-based measure, correlation dimension depends both on the number of variable sites and the amino acid usage, U, the average number of different amino acids found at each site. Correlation dimension allows "stacking" of partially occupied dimension axes of different sites, leading to the estimate which might be smaller than the number of variable sites. This contrasts with the topological sequence dimension (D_{top}), which we define as the number of sites with nonzero amino acid variation, independent of U.

The volume or size of occupied sequence space can be calculated from dimensionality as $V \approx {r_{max}}^{D_{cor}}$, where r_{max} is the maximum distance between two sequences in a given family in amino acids. We derive the relationship $D_{top} \approx log_U(V) = log_U({r_{max}}^{D_{cor}})$linking the estimated correlation dimension to topological sequence dimension under the assumption of uniform usage of U amino acids per site (*SI Appendix, Methods*). To compare the space sizes of different protein families we define effective topological dimension (similar to the notion of "effective" sequence length in ref. 65) which uses the common base of the logarithm 20—for the maximum possible amino acid usage. To demonstrate the logic behind this, consider the following example. A set of sequences of length 11 varying at all sites with an amino acid usage of only five and an effective topological dimension of ~7.5 will occupy the same amount of sequence space (and correspond to the same number of sequences) as a set of sequences of length six having maximal usage of 20 at each site (*SI Appendix*, Fig. S5*B*). Topological sequence dimension (equal to sequence length) for the same example would overestimate the amount of occupied space for the first set by five. Throughout the text we use the effective topological dimension as a more intuitive measure, which we calculate from the correlation dimension (which we report in *SI Appendix*).

Selective constraint reduces the dimensionality of the sequence space. For example, strong purifying selection that allows only a single amino acid at an invariable site reduces the topological dimension of sequence space (66). Similarly, variable sites in which natural selection restricts evolution to accepting fewer than 20 amino acids will have a lower correlation dimension. Thus, the maximum theoretical dimensionality of a protein family strongly depends on α, the fraction of allowed amino acid states at any given time (Fig. 1*B*). Protein families typically have very few invariable sites and many have a relatively high α (67), which can be approximated using the rate of protein evolution (dN/dS > 0.2, *SI Appendix, Methods*). These observations lead to a naive expectation that protein family sequence distribution would have a high correlation dimension. However, for a handful of proteins the dimensionality was estimated to be at least an order of magnitude smaller than the theoretical maximum (23, 55). Here, we expand

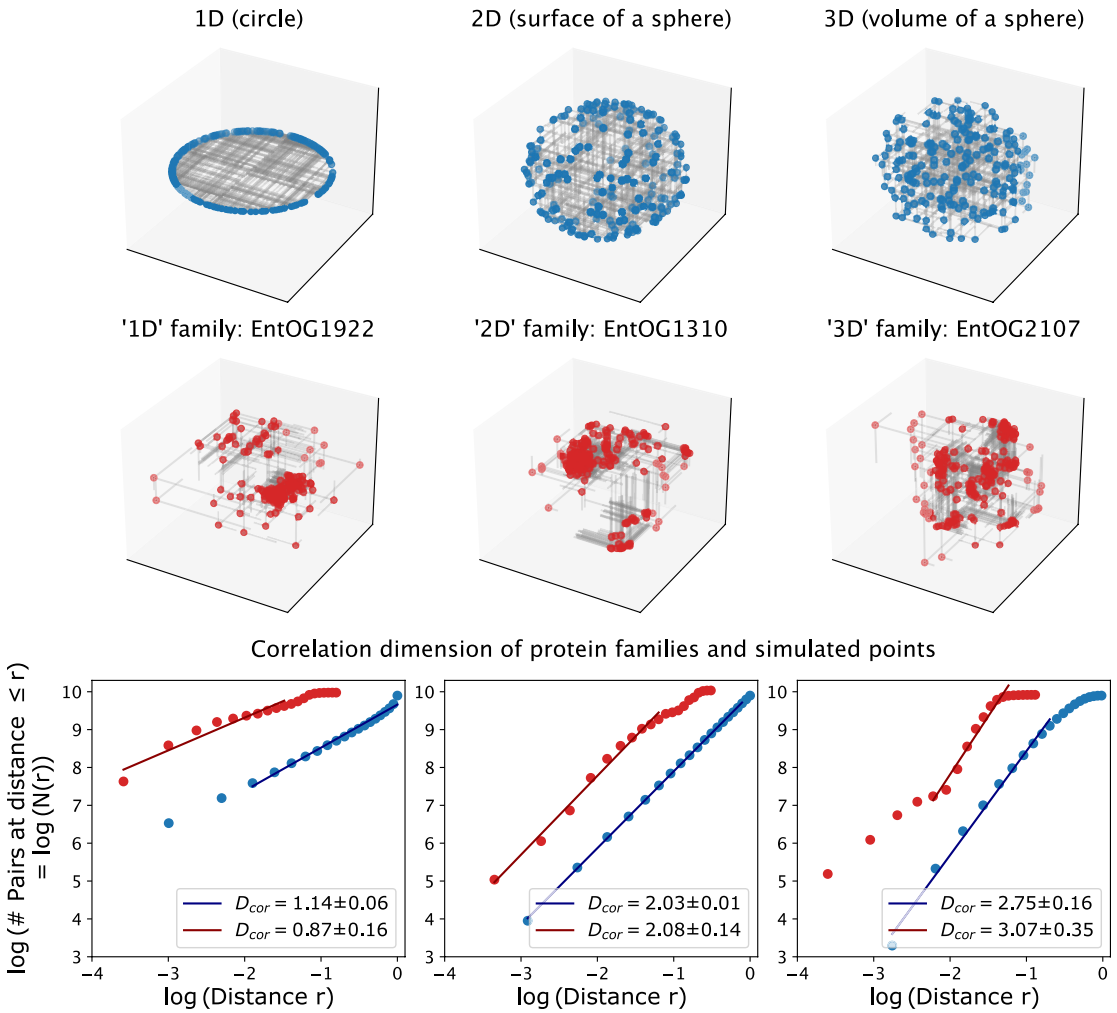

Fig. 2. Correlation dimension on points and protein sequences. *Top*: points simulated in 1D (circle), 2D (surface of the sphere) and 3D (volume of the sphere). *Middle*: projections of the metric multidimensional scaling embeddings of ~200 sequences for three protein families in 3D. *Bottom*: curves used for correlation dimension estimation (also called correlation integrals) for corresponding sets of points from the *Top* (blue) and *Middle* (red) panels and the estimated correlation dimension D_{cor} ± 95% CI (*Insets*, calculated as a slope of the depicted regression line $\log(N(r)) = D_{cor}\log(r) + \text{const}$, *SI Appendix, Methods*). The corresponding effective topological dimensions for shown protein families are 1.27, 3.26, 4.17 (*Left* to *Right*). The *Top* and *Bottom* sections are adapted from ref. 64.

on these studies by undertaking a proteome-wide study of correlation dimension in three selected phylogenetic groups with the aim to understand the biological factors that restrict the distribution of extant sequences in sequence space.

To estimate correlation dimension for a protein gene family it is necessary to calculate distances between homologous sequences in the gene family. In principle, considering homologous proteins from the entire tree of life for each protein family would allow us to study a large area of the sequence space. However, in our experience, large multiple sequence alignments of broader phylogenetic groups are of poor quality. Thus, we settled on Vertebrata, Enterobacterales, and Gammaproteobacteria as the main datasets, because Vertebrata and Enterobacterales had a comparable number of sequences with similar sequence divergence across gene families and also included Gammaproteobacteria to compare with Enterobacterales, controlling for phylogenetic depth of our data.

Using multiple sequence alignments of hundreds of homologous protein families from these three phylogenetic groups we estimated their correlation and effective topological dimension. We found that the estimated dimensionality was much smaller than what may be expected (Fig. 1*B*) for long proteins in all protein families in all three phylogenetic groups (Fig. 3*A*). For some gene families the effective topological dimension was on the order of 1; in just a few families it was larger than 10 (Fig. 3*A*). Such low dimensionality cannot be explained by low sequence divergence because high pairwise sequence divergence is common across multiple gene families (*SI Appendix*, Fig. S8*B*). Gammaproteobacteria, a deeper phylogenetic group with a higher number and more distant sequences, had higher dimensionality across different gene families. However, unlike sequence divergence, the number of sequences does not increase dimensionality when considering evolution along a phylogenetic tree (*SI Appendix*, Fig. S4*A*), thus, the higher dimensionality in Gammaproteobacteria cannot be explained simply by sample size (*SI Appendix*, Fig. S4 and *SI Appendix*).

Evolutionary Factors that Limit the Dimensionality of Explored Sequence Space. Three factors that may cause a small dimensionality of extant sequences have been considered previously. Selective constraint limits the exploration of sequence space (67), so slow-evolving gene families should have lower dimensionality than faster-evolving ones (Fig. 1*B*). Epistasis also prevents the exploration of sections of sequence space (11), potentially reducing dimensionality of extant sequences (23). Finally, extant sequences are not uniformly distributed in sequence space because they share an evolutionary history through a phylogenetic process, likely having a strong effect on their dimensionality (23).

To better understand why the dimensionality of extant sequences is so small across all protein families, we undertook a

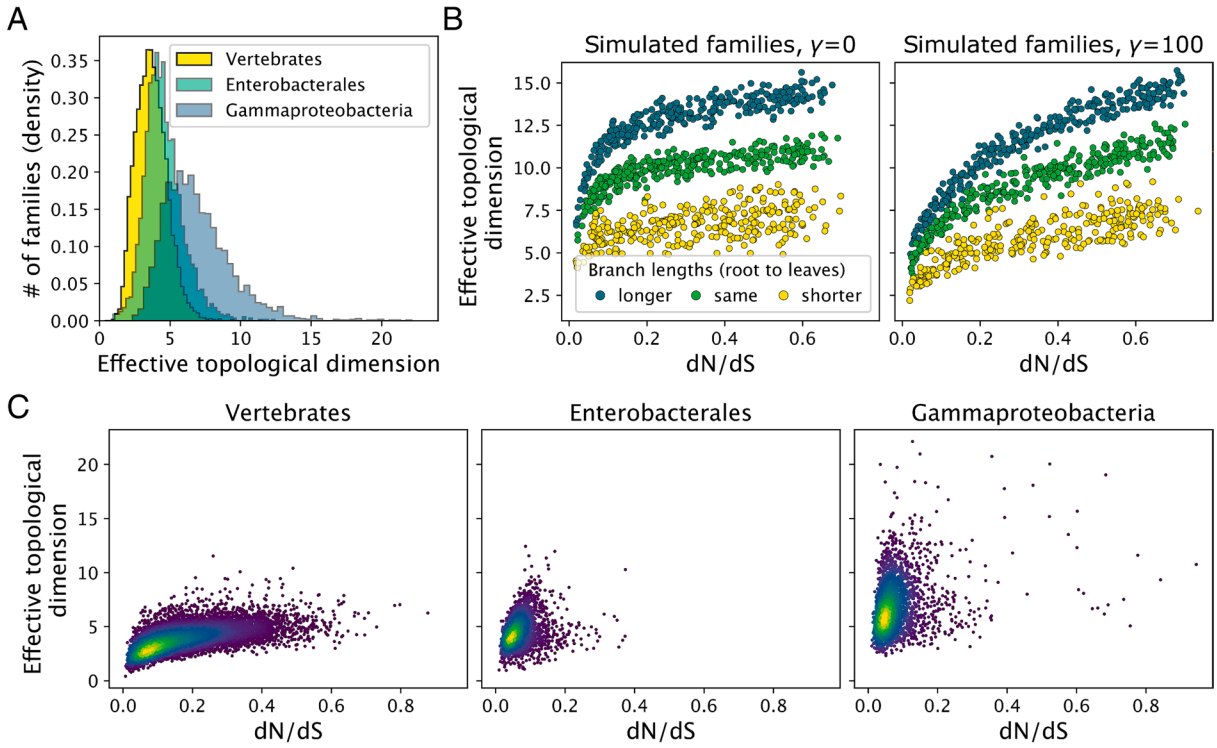

Fig. 3. Dimensionality of natural and simulated sequences. (*A*) The distribution of the effective topological dimension in different protein families. (*B*) Dimensionality of a simulated protein family as a function of mean dN/dS per family (an empirical estimate of α, the fraction of allowed amino acid states), and branch lengths (color). (*C*) Estimated dimensionality as a function of mean dN/dS per family for natural sequences. Brighter colors (more yellow) indicate higher point density.

systematic analysis of how these three factors impact correlation dimension. We simulated sequence evolution under varying extent of selective constraints, different extent of epistasis and different phylogenetic tree shapes. An ideal modeling approach would be to simulate sequence evolution on a predefined, complex, and multidimensional fitness landscape. However, since the number of underlying sequences that are needed to define a fitness landscape is greater than the number of elementary particles in the universe, a different approach is needed. Thus, we modeled protein evolution using the fitness matrix approach (68) that models protein evolution along an evolutionary trajectory with epistasis but without defining an impossibly large fitness landscape. We used the fitness matrix model to estimate the expected dimensionality of simulated protein sequences under a range of realistic parameters. We describe how we did the simulations in detail in *SI Appendix*, *Methods*, *Simulated Protein Families*, but provide a brief description here. In our model fitness is binary, so substitutions are either allowed or not allowed to occur. The model explicitly considered the extent of selective constraint (α) as the proportion of amino acid states allowed at any given moment. The extent of epistasis was modeled by the parameter γ. When γ was 0, there was no epistasis and the fitness matrix was constant so that the same amino acid substitutions were always allowed. However, when epistasis is present, after one nonsynonymous substitution, on average γ allowed substitutions become not allowed and vice versa. We also considered four different phylogenetic tree shapes—a starlike tree, random topology, and trees with branch lengths dependent on their distance from the root (*SI Appendix*, Fig. S9). The simulations resulted in protein families consisting of 300 sequences located at an equivalent pairwise distances as what we observed in real protein families.

When we simulated sequence evolution along a phylogeny the effective topological dimension of these sequences was similar to that in real gene families across all dN/dS values (Fig. 3*B*). Slow-evolving protein families (with low average dN/dS) tended to have a smaller dimension, showing a diminishing returns relationship in vertebrates but not in bacterial phylogenetic groups (Fig. 3*C*). The plateau of dimensionality at high dN/dS values might be due to the fact that more amino acid states are available than are realized during evolution of fast evolving families. Thus, the lack of the diminishing returns relationship in the bacterial clades may reflect the lower dN/dS values for most bacterial families. Crucially, even for gene families with high dN/dS values the effective topological dimension remained orders of magnitude below the theoretical maximum (Fig. 1*B*). Thus, selective constraint may explain why the effective topological dimension of a specific gene family is 1 rather than 2, but it cannot explain why the effective topological dimension is 2 and not 200.

An epistatic fitness landscape may contain impassible fitness valleys and isolated fitness peaks (ref. 19), which is frequently invoked as a factor that can limit evolutionary exploration of the sequence space (18, 69). These features may reduce correlation dimension because the distribution of extant sequences on such a landscape may be more sparse than on a nonepistatic one. Thus, we considered how the extent of epistatic interactions affects the distribution of extant sequences. As expected (34), when epistasis was common, the number of variable sites, amino acid states per site (amino acid usage), and pairwise sequence divergence were higher (Fig. 4*A* and *SI Appendix*, Fig. S10). By contrast, more epistatic interactions between sites reduced the dimensionality of sequences evolving along a phylogeny. The reduction of dimensionality in simulations with epistasis is likely due to clustering of extant sequences from subclades of the phylogenetic tree due to epistatic entrenchment (70, 71) (*SI Appendix*).

It may seem contradictory that epistasis increases amino acid usage (34) and pairwise sequence divergence but does not increase

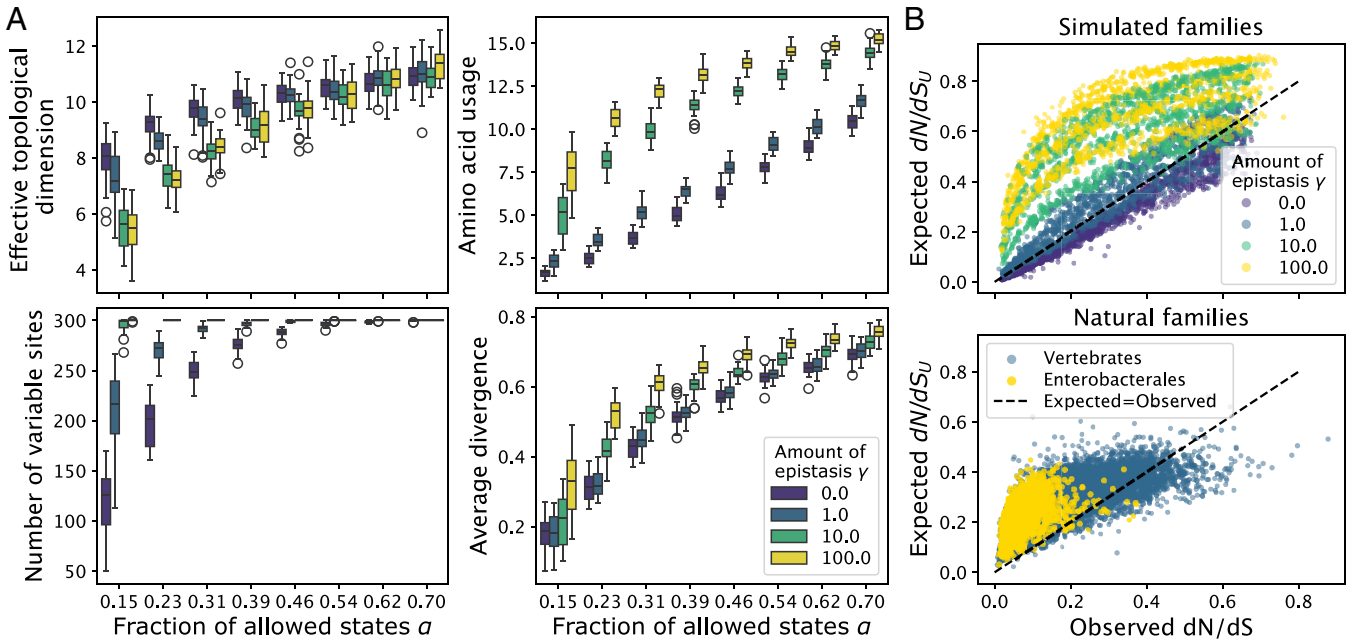

Fig. 4. Effect of epistasis on the exploration of the sequence space and dimensionality of simulated sequences. (*A*) Effective topological dimension, average amino acid usage, number of variable sites and average pairwise distance as a function of α, the fraction of allowed states, and γ, epistasis strength (color). The simulations are for a binary phylogenetic tree with no branch lengths scaling with depth. (*B*) Epistasis in natural and simulated sequences estimated using amino acid usage. *Top*: dN/dS_U in sequences simulated without epistasis (purple) is equal or below the observed dN/dS. *Bottom*: dN/dS_U of natural sequences is mostly above the observed dN/dS, resembling the dN/dS_U of the simulated sequences with epistasis from the *Top* plot.

the dimensionality of the distribution of extant sequences (Fig. 4*A*). This apparent contradiction can be resolved when we consider that epistasis can have different effects at different scales. On a local scale, within the vicinity of an evolving sequence, epistasis may have a restrictive role, preventing substitutions into amino acid states that are not available in the current genetic context (70, 71). However, on a global scale, epistasis allows for the exploration of very distant areas of the sequence space (15, 46, 72), eventually allowing the incorporation of the locally restricted amino acids. Thus, epistasis can make some amino acid states inaccessible in the short term, but in the longer term it will eventually allow evolution to explore a larger amino acid state diversity.

The discrepancy between the local and global availability of amino acid states, as estimated by mean dN/dS and amino acid usage (U), respectively, allows us to quantify the extent of epistasis affecting the analyzed proteins (34) (Fig. 4*B*). Without epistasis the measured mean dN/dS should equal $dN/dS_U = (U-1)/19$, or the fraction of the globally available amino acid states. Even though there is extensive epistasis that shapes evolution of the protein families, within a realistic parameter range of dN/dS and α (Fig. 4*B*), epistasis cannot substantially reduce dimensionality of extant sequences (Fig. 4*A*) (23). Thus, epistasis is a restrictive factor on the local fitness landscape scale but globally it is a permissive one.

Phylogeny has a large effect on dimensionality (23) and we explored the effect of tree shape on gene families in our dataset. We found that simulated evolution on trees in which internal branch lengths increase from root to leaves resulted in higher dimensionality (Fig. 5*A*). Nevertheless, when trees with long terminal branches were used the dimensionality was still much smaller than the theoretical maximum (Figs. 1*B* and 5*B*). We thus considered the tree with longest possible terminal branches—a starlike phylogeny. As expected, evolution on a starlike phylogeny resulted in an even higher dimensionality; however, even then it was much lower than the theoretical limit (Fig. 5*A*).

Indeed, the only way to obtain a dimensionality equal to the theoretical maximum was to randomly sample the functional sequence space that was determined by a nonepistatic fitness matrix (Fig. 5*B*). The difference between a starlike phylogeny and a random exploration of sequence space is that in the former case exploration the sequence space occurs starting from a single point of origin, while in the latter case the sequence space is sampled uniformly across the space without the restriction by a single point of origin. Thus, the main factor that limits dimensionality of extant sequence space is that sequence evolution of a gene family starts from a single point of origin, which is the common ancestor sequence.

Scaling the Point of Origin Effect. Sequence evolution from a single point of origin limits evolutionary exploration of the sequence space to an extreme extent. Phylogenetic tree topology has the second strongest effect, in part because the phylogenetic nature of evolution is effectively a series of local point of origin effects: each branching point on the tree is a new point of origin limitation for subsequent evolution. Because of the point of origin effects, when two homologous sequences are aligned to measure correlation dimension, a substantial fraction of sites will have identical states by descent from a common ancestor (Fig. 6*A*). By contrast random sequences sampled from the entire nonepistatic fitness landscape are combinations of all allowed amino acid states, which were never tested by evolution because these combinations are many substitutions away from any sequences that ever existed. Thus, two such randomly sampled sequences will be mostly different from each other, except for rare instances when the same amino acid was independently chosen in the same site (Fig. 6*A*). Indels and recombination will be subject to the same point of origin effects and are also not expected to substantively contribute to deep exploration of protein sequence space.

The extent to which the point of origin effect limits exploration of protein sequence space is best illustrated with an example. Assuming constant α and γ, we can use dN/dS to derive the upper

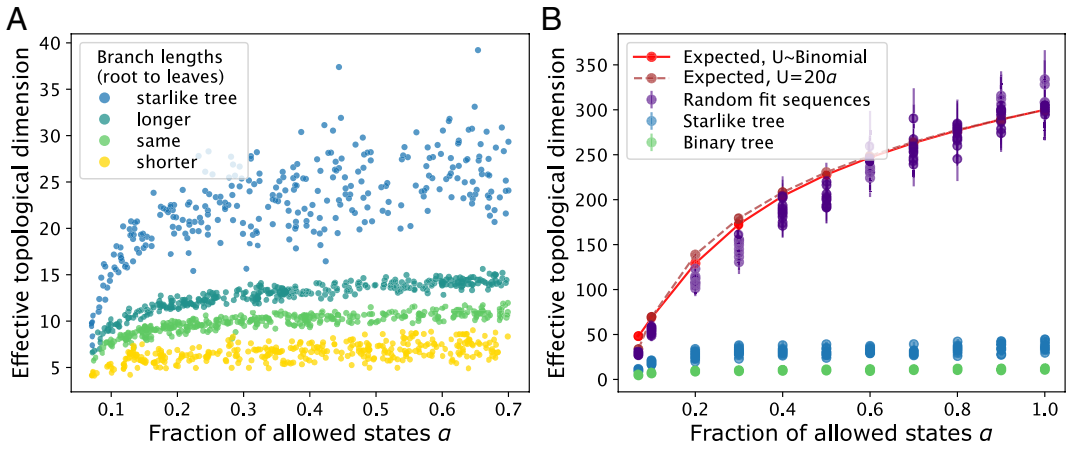

Fig. 5. Impact of phylogenetic tree topology on dimensionality. (*A*) Effective topological dimension of protein families as a function of α, fraction of allowed amino acid states, with no epistasis (γ = 0). We simulated different tree topologies while maintaining the same distribution of the pairwise sequence divergence among simulated sequences. (*B*) Simulated evolution of a protein family from a single point of origin considerably decreases dimensionality, while the dimension of randomly sampled sequences matches the expectation (from Fig. 1*B*). We simulated protein families with 1,000 sequences 300 amino acids long each and no epistasis. Ten replicates per parameter set are shown, error bars represent 95% CI.

limit of the functional volume of sequence space that takes into account the limitation of selection (α) and epistasis (γ). For the DnaA protein, we estimate that the hypothetical upper limit of the functional volume of sequence space has 10^{36} more sequences than the extant sequence space of its Enterobacterales homologues. Assuming constant α and γ is unrealistic and leads to an overestimate of the volume of functional sequence space. However, 10^{36} is such an astronomically large volume of sequence space that it seems certain that among these sequences is a very large number of functional DnaA sequences that were unexplored in evolution due to the point of origin effect. DnaA is not exceptional, and the differences between the observed extant volume of sequence space and the dN/dS-derived hypothetical maximum is many orders of magnitude across most gene families (Fig. 6*D* and *SI Appendix*, Fig. S11).

How much the point of origin effect restricts exploration of the sequence spaces depends on how long sequences evolve from their common ancestor. With sufficient time the difference between random sampling of a fully connected ridge on a fitness landscape and its exploration through an evolutionary phylogenetic process disappears (Fig. 6*B*). Thus, the reason why the dimensionality of extant sequences from the older Gammaproteobacteria was higher than of its younger Enterobacterales subgroup (Fig. 3*A*) is likely because Gammaproteobacteria diverged for a much longer time. The common ancestor of Gammaproteobacteria existed ~1.5 Bya, and across all protein families the effective topological dimension of extant sequences was from five to eight (Fig. 3*A*). Evolution since last universal common ancestor (LUCA) proceeded only about 2.5 × longer than evolution since the common ancestor of Gammaproteobacteria, correspondingly, the effective topological dimension taken from a prokaryote-wide sequence dataset [Clusters of Orthologous Genes (74), *SI Appendix*, *Methods*] was proportionally larger, from 10 to 16 for most protein families (*SI Appendix*, Fig. S8*A*). Thus, Gammaproteobacteria explored, on average, only 10^{-64} (10^{-338} under nonepistatic scenario) of the hypothetical functional sequence space, while prokaryotes, in the course of about 4 billion years of evolution, explored only about eight orders of magnitude more (Fig. 6*D*). The number of substitutions per site since LUCA is thought to be ~1 (75) and sufficiently high dimensionality of evolving sequences can be reached when the length of the tree reaches an average of ~9 amino acid substitutions per site (Fig. 6*B*). Thus, looking into the distant future, protein families are unlikely to fully explore the functional sequence before stellar evolution of the Sun eliminates life on Earth altogether.

Limitations and Assumptions. The point of origin effect has a markedly stronger influence on the dimensionality of extant sequences than the other factors we considered. However, the reported dimensionality values for each extant or simulated protein family are not exact. Indeed, the underlying data are subjected to a variety of influences entrenched in data selection and analysis. On a fundamental level, accuracy of the correlation dimension estimate can be influenced by just three factors: the number of points, bias of their sampling or by inaccuracies in measuring distances between them. Because sequence space is highly dimensional, a large number of points in space is needed for an accurate estimate of correlation dimension (63). Thus, at first glance, it seems that the best scenario for accurate dimension estimates would be to use all available orthologs per gene family. However, many thousands of highly diverged sequences are hard to align accurately (76) due to the large number of indels that may ultimately distort the measured distances between sequences. Selection of a specific phylogenetic group for such an analysis can also introduce biases, including overrepresentation of sequences from culturable and model species, or nonrandom sampling of sequence space due to evolutionary entrenchment (71). Our choice of using Gammaproteobacteria and vertebrates is, therefore, a compromise between size, sampling bias, and phylogenetic depth created to include enough sequences for estimating dimensionality while maintaining feasibility of the needed computational time to calculate pairwise distances between a large number of sequences. Even though the exact datasets used for most of the analysis cover only certain phylogenetic groups, because of the similarity of our results for such dissimilar taxons (vertebrates and Bacteria) we believe they can be generalized. Finally, we used the Clusters of Orthologous Genes (COGs) dataset (74) of all prokaryotes to get an idea of the maximum attainable dimensionalities.

Estimated dimensionality of simulated sequences is subject to the same three factors, which guided our modeling approach. We simulated a similar number of sequences to that observed in our extant dataset to be able to compare the two (see *SI Appendix*, Fig. S4*A* for results of simulation of a large number of sequences). We simulated evolution on binary tree topology, which did not

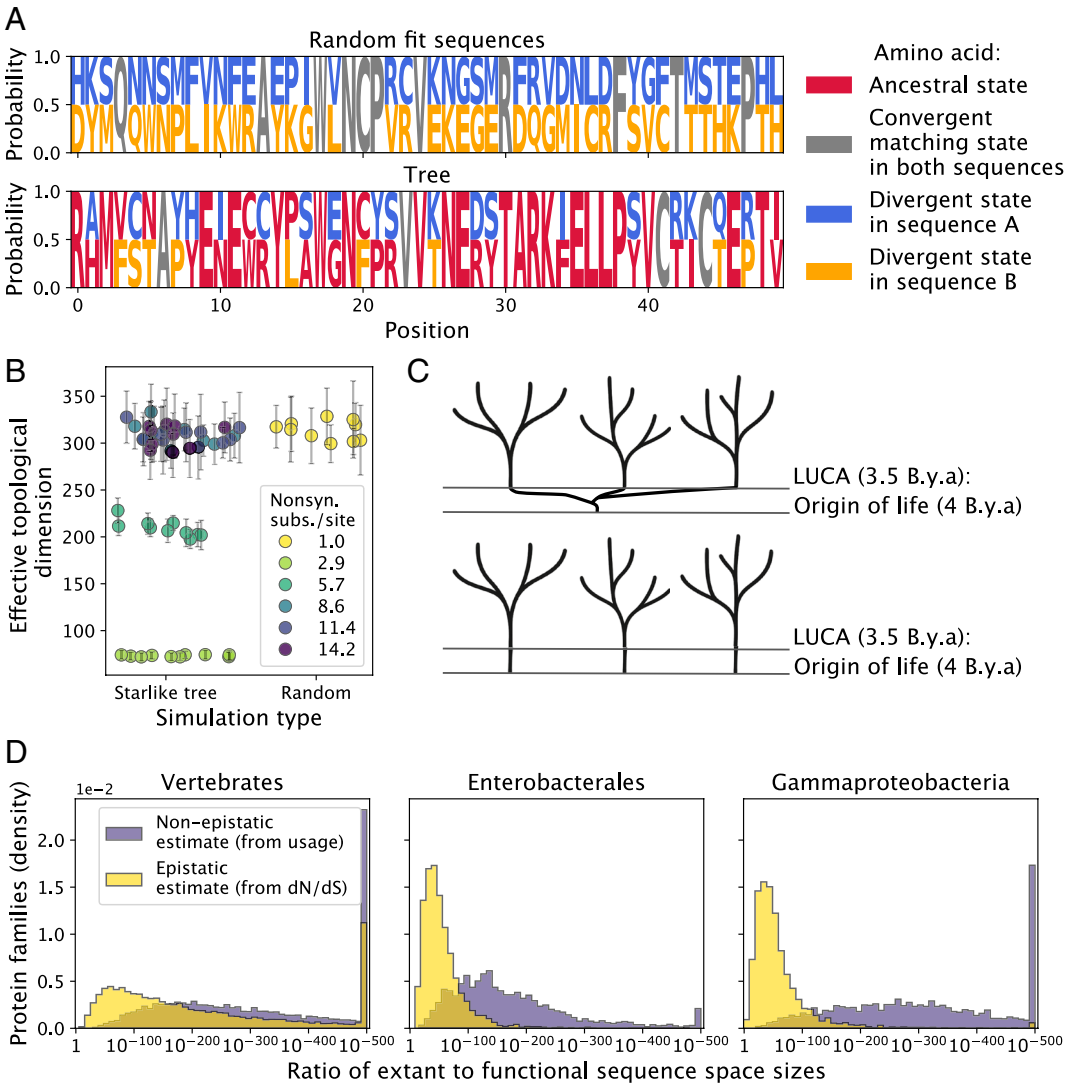

Fig. 6. Implications of the point of origin effect. (*A*) Alignment of two sequences when they do not share (*Top*) and share (*Bottom*) a common ancestor. (*B*) Effective topological dimension of families with 1,000 sequences evolving under neutrality on a starlike phylogenetic tree as a function of time (average number of substitutions at a site shown in color) compared to the dimensionality of randomly sampled fit sequences. Whiskers show 95% CI. (*C*) Scenarios of origin of different protein families from a common pre-LUCA ancestor (*Top*) or independent evolution (*Bottom*). (*D*) Fraction of functional sequence space occupied by extant proteins as estimated without epistasis (purple distribution) calculated from amino acid usage as in ref. 73 and under epistasis (yellow distribution) calculated using mean dN/dS for each family. Values equal to or below 10^{-500} are grouped together.

take into account extinction and horizontal transfer events. The simulated protein families thus represent a simplified view of sequence evolution, possibly influencing correlation dimension estimates. We used dN/dS to estimate the strength of selective constraint and the expected dimension and volume of the functional sequence space. These estimates will be biased by positive selection (77) and be more noisy in short genes (78). Further, dN/dS measurements in large clades can bias average estimates as the fraction pairs satisfying dS < 0.8 cutoff decreases with evolutionary distance.

Perhaps the most important limitation of our work is how we model the global fitness landscape. We use the fitness matrix model, which is a good approach to study the evolution of sequences along a phylogeny, but it has inherent assumptions that bias its representation of the entire sequence space. This model assumes that γ and α are constant, which may hold for sequences located on a phylogeny, but is clearly not true across larger regions of the sequence space—the probability that a sequence is nonfunctional is much lower in the vicinity of a wildtype (WT) sequence than in the vicinity of a random sequence (26, 41). Some parameter of the rate of change of α (and γ) as a function of distance may make the fitness model applicable across wider areas of the sequence space. The study of how α and γ changes across sequence space is in its infancy (40) and remains, in our opinion, the most interesting question about the global nature of the fitness landscape.

Discussion

The inefficiency of exploring functional sequences space through a phylogenetic process due to the point of origin effects has interesting implications about the distant evolutionary past. How did molecular evolution lead to the protein sequences that were found in LUCA? Our data support the hypothesis that a substantial proportion of protein functional diversity found in LUCA must have appeared from independent evolution. We have shown that within each protein family extant sequences populated only a very small area of the sequence space over the

course of the last ~3.5 billion years of evolution. It seems highly improbable that in the relatively short period of 0.5 billion years of evolution preceding LUCA evolution would have been able to explore a much wider area of the sequence space and have created a multitude of different protein functions from a single point of origin. A much more likely scenario appears that most of the protein functional and sequence diversity found in LUCA do not share a common ancestor (Fig. 6*C*). These data lend independent support to studies suggesting that pre-LUCA protein evolution proceeded not through accumulation of amino acid substitutions in long sequences but through combination of smaller protopeptides into larger ones (79–81). Similarly, a substantial proportion of clade-specific protein families may have also evolved de novo rather than by virtue of common descent from existing protein sequences (82, 83).

The argument that out of 10^{36} hypothetically functional sequences of DnaA a large fraction must actually be functional but have not been explored by evolution is an adaptation of the argument for the existence of nonhuman intelligent life in the Universe because it is likely given the estimated 10^{18} planets in existence. We can extend this analogy to the Drake equation and derive a simple equation that estimates the probability of emergence of de novo functional protein sequences. The number of useful functional de novo sequences that emerge on Earth per unit time (generation or germline cell division), N, is the number of all existing genomes, G_e, multiplied by the rate of emergence of expressed de novo protein sequences, R_t, multiplied by the fraction of those that are functional, f_f, multiplied by the fraction of those that are selected for in the specific genome in which they emerged, f_s.

$$N = G_e * R_t * f_f * f_s. \quad [1]$$

Unfortunately, the values of these parameters are extraordinarily difficult to estimate. Nevertheless, a back of the envelope calculation can be instructive. G_e has been estimated as ~10^{30} (84, 85) while it is harder to estimate R_t. One R_t estimate can be derived by taking estimates of de novo transcription of ORFs. On the order of 1,000 de novo genes with protein coding potential were found in *Drosophila melanogaster* (86), which may have a global census population size of ~ 10^{13}, making $R_t \approx 1{,}000/10^{13} \approx 10^{-10}$. Another way to obtain a rough estimate is to consider that the probability that a new organism expresses a sequence not found in its parents must be substantially lower than the rate of mutation, or < 10^{-9}. Both these approaches suggest $10^{-10} > R_t > 10^{-15}$, implying that functionally useful protein sequences of any length emerge de novo each generation on Earth if $f_f * f_s$ is > 10^{-15} ($G_e * R_t / f_f * f_s > 1$). Thus, throughout Earth's history if $f_f * f_s \approx 1$, approximately 10^{11} functionally useful de novo sequences emerged (~10^9 y of evolution with 100 generations per year), out of which only ~1% would have survived genetic drift (87). This estimate appears to be broadly consistent with independent observation that de novo genes appear at a modest rate in a number of different lineages (88–90).

Taking aside the obvious issue of the accuracy of the order of magnitude estimates of G_e and especially R_t, the fraction of functional sequences among random ones, f_f, appears to be the central question to understanding the nature of protein sequence space. Only one comprehensive empirical estimate was performed for ATP-binding function (61, 91, 92), measuring $f_f \sim 10^{-11}$ for a protein segment 80 amino acids long. For 18 proteins a rough computational attempt to estimate the number of functional sequences have been made, ranging from 10^9 to 10^{122} for various sequences 35 to 153 amino acids long (26, 49, 93). Although direct modeling of functional proteins may be used to estimate f_f in some circumstances (94), direct empirical data are desired as much as they are lacking. The difficulty is that if the sequence space for many protein functions may be very sparse—if 10^{-100} random sequences are functional, it would be impossible to find even a single functional sequence among randomly synthesized ones. Thus, at present such an approach may be feasible only for relatively short sequences in a fortuitously permissive functional context such as ATP-binding (91). Furthermore, the number of possible sequences grows exponentially with sequence length, thus, f_f likely decreases substantially with length (see table 2 in ref. 26).

The practical implication of our results is that extant or ancestral sequences do not fully describe the entire functional sequence space. Thus, predictions of functional proteins based on extant protein sequences or structures potentially miss a vast majority of all possible useful protein functions. AI-based approaches for "deep" sequence space exploration look promising (5, 7, 95), however, they have been successful mostly for relatively short proteins that likely have a large f_f. A key existing limitation of these types of studies is that experimental validation often consists of verification of expression, solubility, and folding, but not the biological function. When functions are experimentally tested they include mainly binding other molecules (proteins, nucleic acids, etc.) (5, 95), but even several examples of de novo enzyme design proved successful (6, 96). Presently, truly de novo design of larger proteins with complex functions (such as enzymes) with low or even undetectable similarity to known sequences remains a challenge (see ref. 1 for a more complete review). We suggest that an experimental exploration of sequence space that has not been explored in evolution is a necessary component to train the new generation AI models for successful de novo protein design.

The vastness of the evolutionarily unexplored functional sequence space suggests that many novel protein functions are still to be discovered by evolution or by protein design. Researchers interested in designing new proteins are likely restricted to starting from known functional protein sequences (either natural or designed and experimentally validated) and encroaching into the evolutionarily unexplored territory (1, 92). One approach to explore such novel territory of the sequence space is to preferentially incorporate amino acid states that have not been observed in extant or ancestral proteins (40, 56). Another approach is a stepping stone empirical exploration, whereby a local fitness landscape of a WT sequence is used to predict a functional protein in an unexplored area of the fitness landscape (8), then another local fitness landscape is created for this novel protein that predicts an even more distant protein, ad infinitum. Thus, the key to understanding the global nature of the fitness landscapes will be to accumulate information on the functional artificial proteins far outside the functional sequence space that has been explored by evolution.

Methods

We used homologous sequence alignments from several major phylogenetic groups, Vertebrata, Enterobacterales, Gammaproteobacteria, and Prokaryota (COGs). For each protein family in each group, we measured all pairwise distances between homologues, estimated the correlation dimension, amino acid usage, and mean pairwise dN/dS from among all pairs of sequences with dS < 0.8. We performed simulations of protein evolution using vertebrate gene trees as a model template of branch length distribution. The single simulated gene sequences after evolution on a phylogeny were subject to the same measurements of pairwise distance, correlation dimension, amino acid usage, and mean dN/dS.

Data, Materials, and Software Availability. All of our code used to obtain all figures: https://github.com/oist/sequence-space-dimension (97). Data are available at Figshare: (98).

ACKNOWLEDGMENTS. We thank Olga Kalinina for feedback on our manuscript, Vsevolod Kuksin for fruitful discussions and Lev Tsarin for participation in the design of our models. This work was supported by Japan Science and Technology Agency as part of Adopting Sustainable Partnerships for Innovative Research Ecosystem, Grant No. JPMJAP24B2 (F.A.K. and L.H.I.), and Fonds Zur Förderung der Wissenschaftlichen Forschung Grant ESP253-B (O.O.B.)

Author affiliations: [a]Evolutionary and Synthetic Biology Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan; [b]Institute of Science and Technology Austria, Klosterneuburg 3400, Austria; [c]Centre for Microbiology and Environmental Systems Science, Department of Microbiology and Ecosystem Science, Division of Computational System Biology, University of Vienna, Wien 1030, Austria; and [d]Centro de Astrobiología, CSIC-INTA, Torrejón de Ardoz, Madrid 28850, Spain

1. J. Yang, F.-Z. Li, Y. Long, F. H. Arnold, Illuminating the universe of enzyme catalysis in the era of artificial intelligence. *Cell Syst.* **17**, 101372 (2025).
2. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
4. J. Meier *et al.*, Language models enable zero-shot prediction of the effects of mutations on protein function. bioRxiv [Preprint] (2021), https://doi.org/10.1101/2021.07.09.450648. Accessed 29 October 2025.
5. A. M. Subramanian, Z. A. Martinez, A. L. Lourenço, S. Liu, M. Thomson, Unexplored regions of the protein sequence-structure map revealed at scale by a library of foldtuned language models. bioRxiv [Preprint] (2025), https://doi.org/10.1101/2023.12.22.573145. Accessed 19 May 2025.
6. A.H.-W. Yeh *et al.*, De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
7. C. R. Freschlin, S. A. Fahlberg, P. Heinzelman, P. A. Romero, Neural network extrapolation to distant regions of the protein fitness landscape. *Nat. Commun.* **15**, 6405 (2024).
8. T. Hayes *et al.*, Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
9. T. C. B. McLeish, Are there ergodic limits to evolution? Ergodic exploration of genome space and convergence. *Interface Focus* **5**, 20150041 (2015).
10. S. Wright, "The roles of mutation, inbreeding, crossbreeding and selection in evolution" in *The Proceedings of the Sixth International Congress of Genetics* (1932), vol. **1**, pp. 356–366.
11. J. A. G. M. de Visser, J. Krug, Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
12. A. Heger, L. Holm, Towards a covering set of protein family profiles. *Prog. Biophys. Mol. Biol.* **73**, 321–337 (2000).
13. J. Z. Chen, B. Gall, S. B. Pulsford, N. Tokuriki, C. J. Jackson, Exploring large protein sequence space through homology- and representation-based hierarchical clustering. *Mol. Biol. Evol.* **42**, msaf136 (2025).
14. P. C. F. Buchholz, B. van Loo, B. D. G. Eenink, E. Bornberg-Bauer, J. Pleiss, Ancestral sequences of a large promiscuous enzyme family correspond to bridges in sequence space in a network representation. *J. R. Soc. Interface* **18**, 20210389 (2021).
15. J. Maynard Smith, Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).
16. T. Bitard-Feildel, Navigating the amino acid sequence space between functional proteins using a deep learning framework. *PeerJ Comput. Sci.* **7**, e684 (2021).
17. S. Manrubia *et al.*, From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Phys. Life Rev.* **38**, 55–106 (2021).
18. I. Fragata, A. Blanckaert, M. A. Dias Louro, D. A. Liberles, C. Bank, Evolution in the light of fitness landscape theory. *Trends Ecol. Evol.* **34**, 69–82 (2019).
19. S. F. Greenbury, A. A. Louis, S. E. Ahnert, The structure of genotype-phenotype maps makes fitness landscapes navigable. *Nat. Ecol. Evol.* **6**, 1742–1752 (2022).
20. L. Bromham, Why do species vary in their rate of molecular evolution? *Biol. Lett.* **5**, 401–404 (2009).
21. J. Echave, S. J. Spielman, C. O. Wilke, Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17**, 109–121 (2016).
22. P. C. F. Buchholz, S. Fademrecht, J. Pleiss, Percolation in protein sequence space. *PLoS ONE* **12**, e0189646 (2017).
23. E. Facco, A. Pagnani, E. T. Russo, A. Laio, The intrinsic dimension of protein sequence evolution. *PLoS Comput. Biol.* **15**, e1006767 (2019).
24. J. Marchi *et al.*, Size and structure of the sequence space of repeat proteins. *PLoS Comput. Biol.* **15**, e1007282 (2019).
25. T. Senoner *et al.*, ProtSpace: A tool for visualizing protein space. *J. Mol. Biol.* **437**, 168940 (2025).
26. B. J. Miller, A percolation theory analysis of continuous functional paths in protein sequence space affirms previous insights on the optimization of proteins for adaptability. *PLoS ONE* **19**, e0314929 (2024).
27. D. T. F. Dryden, A. R. Thomson, J. H. White, How much of protein sequence space has been explored by life on earth? *J. R. Soc. Interface* **5**, 953–956 (2008).
28. F. A. Kondrashov, A. S. Kondrashov, Multidimensional epistasis and the disadvantage of sex. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12089–12092 (2001).
29. P. C. Phillips, Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).
30. J. A. G. M. de Visser, T. F. Cooper, S. F. Elena, The causes of epistasis. *Proc. Biol. Sci.* **278**, 3617–3624 (2011).
31. D. M. Weinreich, R. A. Watson, L. Chao, Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**, 1165–1174 (2005).
32. C. B. Ogbunugafor, A reflection on 50 years of John Maynard smith's "Protein Space". *Genetics* **214**, 749–754 (2020).
33. F. H. Arnold, Innovation by evolution: Bringing new chemistry to life (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **58**, 14420–14426 (2019).
34. M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, F. A. Kondrashov, Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535–538 (2012).
35. N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, R. Sun, Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, e16965 (2016).
36. F. J. Poelwijk, S. Tănase-Nicola, D. J. Kiviet, S. J. Tans, Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J. Theor. Biol.* **272**, 141–144 (2011).
37. D. J. Kvitek, G. Sherlock, Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet.* **7**, e1002056 (2011).
38. S. Gavrilets, Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12**, 307–312 (1997).
39. V. O. Pokusaeva *et al.*, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet.* **15**, e1008079 (2019).
40. L. Gonzalez Somermeyer *et al.*, Heterogeneity of the GFP fitness landscape and data-driven protein design. *eLife* **11**, e75842 (2022).
41. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
42. D. H. Bryant *et al.*, Deep diversification of an AAV capsid protein by machine learning. *Nat. Biotechnol.* **39**, 691–696 (2021).
43. C. Bank, Epistasis and adaptation on fitness landscapes. *Annu. Rev. Ecol. Evol. Syst.* **53**, 457–479 (2022).
44. S. D'Costa, E. C. Hinds, C. R. Freschlin, H. Song, P. A. Romero, Inferring protein fitness landscapes from laboratory evolution experiments. *PLoS Comput. Biol.* **19**, e1010956 (2023).
45. P. A. Romero, F. H. Arnold, Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
46. I. S. Povolotskaya, F. A. Kondrashov, Sequence space and the ongoing expansion of the protein universe. *Nature* **465**, 922–926 (2010).
47. E. Ferrada, A. Wagner, Evolutionary innovations and the organization of protein functions in genotype space. *PLoS ONE* **5**, e14172 (2010).
48. M. Bisardi, J. Rodriguez-Rivas, F. Zamponi, M. Weigt, Modeling sequence-space exploration and emergence of epistatic signals in protein evolution. *Mol. Biol. Evol.* **39**, msab321 (2022).
49. J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, M. Weigt, Efficient generative modeling of protein sequences using simple autoregressive models. *Nat. Commun.* **12**, 5800 (2021).
50. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 67–72 (2009).
51. X. Ding, Z. Zou, C. L. Brooks Iii, Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **10**, 5644 (2019).
52. D. Sgarbossa, U. Lupo, A.-F. Bitbol, Generative power of a protein language model trained on multiple sequence alignments. *eLife* **12**, e79854 (2023).
53. Q. Du, H. Wang, B. Jiang, X. Wang, Advancing genetic engineering with active learning: Theory, implementations and potential opportunities. *Brief. Bioinf.* **26**, bbaf286 (2025).
54. H. Kemble, P. Nghe, O. Tenaillon, Recent insights into the genotype-phenotype relationship from massively parallel genetic assays. *Evol. Appl.* **12**, 1721–1742 (2019).
55. P. C. F. Buchholz, C. Zeil, J. Pleiss, The scale-free nature of protein sequence space. *PLoS ONE* **13**, e0200815 (2018).
56. S. Gelman *et al.*, Biophysics-based protein language models for protein engineering. *Nat. Methods* **22**, 1868–1879 (2025).
57. W. P. Russ *et al.*, An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).
58. Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, F. H. Arnold, Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8852–8858 (2019).
59. B. J. Wittmann, Y. Yue, F. H. Arnold, Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Syst.* **12**, 1026.e7–1045.e7 (2021).
60. H. Sahakyan, S. G. Babajanyan, Y. I. Wolf, E. V. Koonin, In silico evolution of globular protein folds from random sequences. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2509015122 (2025).
61. C. L. Tong, K.-H. Lee, B. Seelig, De novo proteins from random sequences through in vitro evolution. *Curr. Opin. Struct. Biol.* **68**, 129–134 (2021).
62. P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors. *Physica D* **9**, 189–208 (1983).
63. J.-P. Eckmann, T. Tlusty, Dimensional reduction in complex living systems: Where, why, and how. *Bioessays* **43**, e2100062 (2021).
64. J. Wang, J. Shan, Segmentation of LIDAR point clouds for building extraction (2009), https://www.asprs.org/a/publications/proceedings/baltimore09/0101.pdf. Accessed 11 September 2025.
65. A. A. Lavin, J. Rivas-Santisteban, Limitations of sequence dissimilarity as a predictor of prokaryotic lineage. *Open Biol.* **15**, 240302 (2025).
66. D. A. Kondrashov, F. A. Kondrashov, Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).
67. D. Graur, W.-H. Li, *Fundamentals of Molecular Evolution* (Oxford University Press, ed. 2, 1999).
68. D. R. Usmanova, L. Ferretti, I. S. Povolotskaya, P. K. Vlasov, F. A. Kondrashov, A model of substitution trajectories in sequence space and long-term protein evolution. *Mol. Biol. Evol.* **32**, 542–554 (2015).
69. L. Ferretti, D. Weinreich, F. Tajima, G. Achaz, Evolutionary constraints in fitness landscapes. *Heredity (Edinb.)* **121**, 466–481 (2018).
70. A. V. Stolyarova *et al.*, Senescence and entrenchment in evolution of amino acid sites. *Nat. Commun.* **11**, 4603 (2020).
71. P. Shah, D. M. McCandlish, J. B. Plotkin, Contingency and entrenchment in protein evolution under purifying selection. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E3226–E3235 (2015).
72. A. S. Kondrashov, I. S. Povolotskaya, D. N. Ivankov, F. A. Kondrashov, Rate of sequence divergence under constant selection. *Biol. Direct* **5**, 5 (2010).

73. K. A. Armstrong, B. Tidor, Computationally mapping sequence space to understand evolutionary protein engineering. *Biotechnol. Prog.* **24**, 62–73 (2008).
74. M. Y. Galperin *et al.*, Cog database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
75. I. K. Jordan *et al.*, A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**, 633–638 (2005).
76. E. Garriga *et al.*, Large multiple sequence alignments with a root-to-leaf regressive method. *Nat. Biotechnol.* **37**, 1466–1470 (2019).
77. J. H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
78. W.-H. Li, *Molecular Evolution* (Sinauer Associates, 2007).
79. V. Alva, J. Söding, A. N. Lupas, A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **4**, e09410 (2015).
80. V. Alva, A. N. Lupas, From ancestral peptides to designed proteins. *Curr. Opin. Struct. Biol.* **48**, 103–109 (2018).
81. Y. A. Muthahari, L. Magnus, P. Laurino, From duplication to fusion: Expanding Dayhoff's model of protein evolution. *Protein Sci.* **34**, e70054 (2025).
82. D. Tautz, T. Domazet-Lošo, The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
83. S. B. Van Oss, A.-R. Carvunis, De novo gene birth. *PLoS Genet.* **15**, e1008160 (2019).
84. A. R. Mushegian, Are there 1031 virus particles on Earth, or more, or fewer? *J. Bacteriol.* **202**, e00052-20 (2020).
85. W. B. Whitman, D. C. Coleman, W. J. Wiebe, Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6578–6583 (1998).
86. A. Grandchamp, P. Czuppon, E. Bornberg-Bauer, Quantification and modeling of turnover dynamics of de novo transcripts in *Drosophila melanogaster*. *Nucleic Acids Res.* **52**, 274–287 (2024).
87. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, NY, 1970).
88. M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, D. J. Begun, Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9935–9939 (2006).
89. E. Bornberg-Bauer, K. Hlouchova, A. Lange, Structure and function of naturally evolved de novo proteins. *Curr. Opin. Struct. Biol.* **68**, 175–183 (2021).
90. S. Xia, J. Chen, D. Arsala, J. J. Emerson, M. Long, Functional innovation through new genes as a general evolutionary process. *Nat. Genet.* **57**, 295–309 (2025).
91. A. D. Keefe, J. W. Szostak, Functional proteins from a random-sequence library. *Nature* **410**, 715–718 (2001).
92. J. A. Ascensao, M. M. Desai, Experimental evolution in an era of molecular manipulation. *Nat. Rev. Genet.* **27**, 81–95 (2025).
93. J. A. Vila, About the protein space vastness. *Protein J.* **39**, 472–475 (2020).
94. P. Koehl, M. Levitt, Protein topology and stability define the space of allowed sequences. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 1280–1285 (2002).
95. R. Verkuil *et al.*, Language models generalize beyond natural proteins. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.12.21.521521. Accessed 18 September 2025.
96. A. Madani *et al.*, Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
97. L. Isakova *et al.*, Sequence-space-dimension code repository. GitHub. https://github.com/oist/sequence-space-dimension. Accessed 18 March 2026.
98. L. Isakova *et al.*, Data for "Descent from a common ancestor restricts exploration of protein sequence space". Figshare. https://doi.org/10.6084/m9.figshare.31132711. Deposited 23 January 2026.