Received: 13 May 2021 | Revised: 19 July 2021 | Accepted: 26 July 2021

DOI: 10.1002/pro.4161

THE PROTEIN SOCIETY WILEY

REVIEW

Shifts in amino acid preferences as proteins evolve: A synthesis of experimental and theoretical work

Noor Youssef[1] | **Edward Susko**[2] | **Andrew J. Roger**[3] | **Joseph P. Bielawski**[1,2]

[1]Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

[2]Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

[3]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Correspondence
Noor Youssef, Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada.
Email: nooryoussef03@gmail.com

Abstract

Amino acid preferences vary across sites and time. While variation across sites is widely accepted, the extent and frequency of temporal shifts are contentious. Our understanding of the drivers of amino acid preference change is incomplete: To what extent are temporal shifts driven by adaptive versus nonadaptive evolutionary processes? We review phenomena that cause preferences to vary (e.g., evolutionary Stokes shift, contingency, and entrenchment) and clarify how they differ. To determine the extent and prevalence of shifted preferences, we review experimental and theoretical studies. Analyses of natural sequence alignments often detect decreases in homoplasy (convergence and reversions) rates, and variation in replacement rates with time—signals that are consistent with temporally changing preferences. While approaches inferring shifts in preferences from patterns in natural alignments are valuable, they are indirect since multiple mechanisms (both adaptive and nonadaptive) could lead to the observed signal. Alternatively, site-directed mutagenesis experiments allow for a more direct assessment of shifted preferences. They corroborate evidence from multiple sequence alignments, revealing that the preference for an amino acid at a site varies depending on the background sequence. However, shifts in preferences are usually minor in magnitude and sites with significantly shifted preferences are low in frequency. The small yet consistent perturbations in preferences could, nevertheless, jeopardize the accuracy of inference procedures, which assume constant preferences. We conclude by discussing if and how such shifts in preferences might influence widely used time-homogenous inference procedures and potential ways to mitigate such effects.

KEYWORDS

amino acid preferences, contingency, entrenchment, epistasis, protein evolution, site-specific fitness landscapes

1 | INTRODUCTION

Protein evolution is complex, leaving confounding signals in natural sequences. An evolutionary biologist interested in understanding the evolutionary history of a population, species, or protein must investigate these patterns and decipher their likely causes: Is the observed signal evidence of adaptive evolution, or could it have arisen by nonadaptive processes? To address these questions, we must first have a rigorous understanding of the

patterns emerging from the interplay of random genetic drift and selection to maintain protein function, but in the absence of adaptive processes. To this end, we review nonadaptive evolutionary phenomena and their identifiable footprints in natural sequences.

The space of possible protein sequences is vast. For an average-sized protein of length 300, the number of possible sequences (20^{300}) exceeds the number of atoms in the observable universe (10^{82}). This combinatorial explosion prohibits our ability to fully characterize the sequence-to-sequence (S2S) fitness landscape on which a protein evolves. A more tractable approach is to define the fitness landscape at an individual site in the protein. The *site-specific fitness landscape* is fully defined by a vector of length 20 describing the fitness of the mutant protein created by placing each amino acid at the site given a particular background sequence S, where $f^h(S) = \{f_1^h(S),\ldots,f_{20}^h(S)\}$ defines the fitness landscape at a site h.[1] From fitness landscapes, we can estimate *site-specific propensity landscapes*. Propensity can be defined as the expected frequency with which an amino acid occurs at a site,[2] or the fraction of sequences at thermodynamic equilibrium carrying that particular mutation.[3] The propensity for an amino acid is related to its fitness by

$$\pi_a^h(S) = \pi_a^{(0)} e^{2N_e f_a^h(S)} \Big/ \sum_x \pi_x^{(0)} e^{2N_e f_x^h(S)} \qquad (1)$$

where N_e is the effective population size and $\pi_a^{(0)}$ is the expected frequency of amino acid a in the absence of selection.[4] In this review, we use the more general term *site-specific preference landscape* to describe the relative preferences for amino acids, based on any of the above definitions. Preference landscapes are often normalized so that the sum of all amino acid preferences is equal to one and are usually represented using a heatmap,[1] a sequence-logo plot,[5] or a barplot[6] (Figure 1).

Proteins evolve with various biophysical and evolutionary constraints on their structures and functions. Such selective constraints manifest as differences in preference landscapes among sites and across time. Spatial, or among-site, variability has been extensively studied revealing commonly observed patterns.[7] Buried sites often prefer hydrophobic residues, while surface sites have a higher affinity for hydrophilic amino acids. In addition, preference landscapes at surface sites are usually more uniform, with many residues having similar preferences, than at buried sites, where only a small number of amino acids have high preferences.[8] Failing to account for such spatial variability can jeopardize the accuracy of inference procedures. As a result, various inference methodologies accommodate differences in frequency profiles across sites.[9] Temporal, or across-time, variability in preference landscapes is comparatively less understood. This has lead to the interpretation of temporal rate shifts as evidence of adaptive evolution[10,11]; however, the role of nonadaptive processes, such as neutral evolution in the presence of epistatic interactions between sites, in changing preferences and rates is gaining appreciation.[3,12,13]

We begin by reviewing various nonadaptive phenomena that give rise to temporal shifts in preferences. Then, we discuss evidence for shifted preferences gleaned through analyses of natural sequence alignments. The observed levels of convergence rates, reversion rates, and replacement rates are broadly consistent with nonadaptive evolution. However, this evidence is inferential and indirect—other mechanisms, which we may not yet appreciate may be the ultimate causes of such signals. To more directly quantify the magnitude and prevalence of shifted landscapes, we discuss results from site-directed mutagenesis experiments. The conclusion from these datasets is that amino acid preferences shift over time. However, nonadaptive shifts are usually minor in magnitude and low in frequency. Nevertheless, such minor yet consistent perturbation in preference landscapes lead to detectable variations in rates across time.[8] We end by discussing the consequence these shifts might have on widely-used inference procedures and potential ways to mitigate their effects.

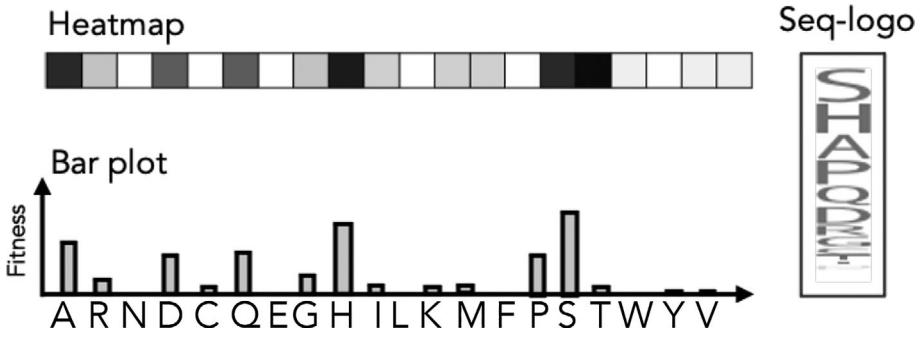


FIGURE 1 Different representations of site-specific preferences. In the heatmap representation, darker shades imply higher preference. In the barplot, bar height represents the preference for the respective amino acid. In the sequence-logo (seq-logo) representation, the size of the letter represents its preference relative to other amino acids

2 | CAUSES OF NONADAPTIVE SHIFTS IN PREFERENCES

Protein evolution is commonly viewed as a walk in sequence space directed by natural selection, drift, and mutations. This was intuitively summarized by John Maynard Smith, where he used a word game as an analogy of protein evolution.[14] Starting with a meaningful word, the objective is to, at each turn, change one letter to yield a different meaningful word. His example trajectory was WORD → WORE → GORE → GONE → GENE. Meaning, in this case, is defined as any English word and is therefore binary (a word is either meaningful or not). Despite its simplicity, Maynard Smith's word game analogy illuminates various salient evolutionary dynamics (Figure 2). Relevant to this review, we will use it to illustrate how adaptive and nonadaptive processes can lead to similar dynamics for site-specific landscapes.

Analogous to a site-specific landscape, let us define a position-specific landscape as a 26-element vector for each letter in the English alphabet. Each letter is assigned a value of zero if it does not produce a meaningful word in the context of the characters present at the other positions and is assigned a value of one otherwise. A change in the background sequence from -ORE to -ONE will cause a shift in the first position fitness landscape. Letters (such as W) that produced meaningful words in the previous background (e.g., WORE) are no longer meaningful in a new background (e.g., WONE). Similarly, letters that were nonviable may become permissible (e.g., DORE versus DONE). In this way, the position-specific landscape is dependent on the background sequence. In proteins, site-specific preference landscapes follow similar dynamics—such context-dependence is referred to as *epistasis*.

It is important to differentiate between shifts in S2S landscapes and shifts in site-specific landscapes. A change in the protein's environment or function will lead to a shift in the ordering of preferred sequences and hence a shift in the S2S landscape. Such a shift is analogous to a change in the definition of a meaningful word (e.g., if Spanish rather than English words are considered meaningful). The evolutionary response to a shift in the S2S landscape is often considered adaptive with an excess of beneficial substitutions compared to neutral or deleterious fixations. Alternatively, site-specific fitness landscapes can change solely due to epistasis in the absence of any external change. In this scenario, the proportions of beneficial and deleterious (fixed by random genetic drift) substitutions remain equal at equilibrium.[15] As such, changes in site-specific landscapes are often considered nonadaptive when the S2S landscape is unchanged. Here, we will refer to *adaptive shifts* as changes in site-specific fitness landscapes in conjunction with a shift in the S2S landscape. Alternatively, *nonadaptive shifts* constitute changes in site-specific landscapes caused by the interplay of mutations, drift, and selection on a fixed S2S landscape.

For most proteins, a prerequisite to proper biological functioning is correct folding into a native structure in which the protein is sufficiently stable. As such, many authors have investigated the level of nonadaptive preference shifts in silico by modeling stability-mediated epistasis and found that amino acid preferences changed over time.[3,8,13] In particular, Pollock et al.[3] observed a tendency for the preference for a resident, recently substituted amino acid to increase through adjustments at other sites in the protein. They refer to this as an *evolutionary Stokes shift*, analogous to the spectroscopy effect known as the Stokes shift in which a molecule receives a

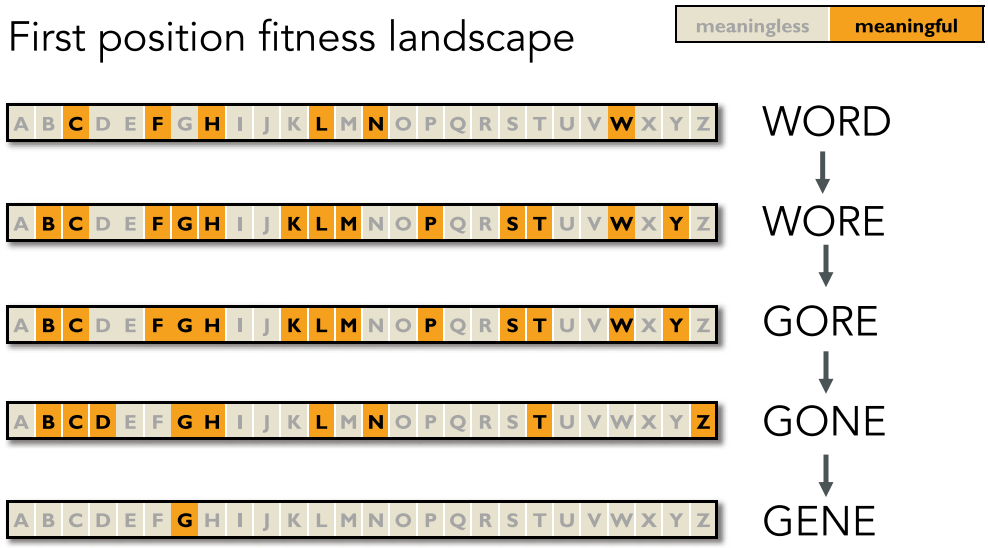


FIGURE 2 Depicting epistatic dynamics using Maynard Smith's word game analogy of protein evolution.[14] The fitness landscape at the first letter position changes as letters at other positions change. Fitness is binary: a word is either meaningful or not. These dynamics are akin to epistatic dynamics in protein evolution where site-specific fitness landscapes depend on the residues present at other sites in the protein

quantum of energy, moves to a higher energy state, and adjusts to the new state by emitting a smaller quantum of energy than was first absorbed. More recently, evidence for the opposite trend, where the preference for the resident amino acid decreases over time, was observed.[2] This phenomenon was dubbed as the *evolutionary anti-Stokes shift*. Using a different stability model, Shah et al.[13] observed similar trends where substitutions were usually *contingent* on prior substitutions that increased their fixation probability, and were subsequently *entrenched*, becoming increasingly deleterious to revert over time.

While entrenchment and evolutionary Stokes shifts are sometimes used interchangeably,[16–18] they are related yet distinct phenomena. Briefly, a substitution may be entrenched "by-any-means" (adaptive or nonadaptive); whereas an evolutionary Stokes shift refers to the increase in preference of a residue by nonadaptive stability-mediated effects. An evolutionary Stokes shift may lead to an entrenched allele; however, not all entrenched alleles result from an evolutionary Stokes shift. Similarly, the notion of contingency and evolutionary anti-Stokes shifts are related yet not synonymous.

To illustrate their differences, consider an adaptive episode where a protein was evolving in the context of Environment A when an external change occurs (Environment B) with a shift in the S2S landscape and accompanying changes in the site-specific landscapes. Let us consider the dynamics at a focal site. In Environment A, amino acid alanine (one-letter code A) was the most preferred residue at a site (Figure 3). In Environment B, the site's preferences change such that valine (one-letter code V) is now the most preferred residue. Assuming that a mutation to a codon specifying V arises at this site, positive selection will then likely lead to its fixation. The substitution to V is therefore contingent on the environmental change that increased its favorability. Once on (or near) the new landscape peak, mutations away from amino acid V will be purged by purifying selection. The beneficial effects of subsequent mutations at other sites may depend on the presence of V as part of the genetic background. As such, substitution away from V may become increasingly deleterious, leading to a degree of entrenchment. In this way, a residue may be contingent and subsequently entrenched through an adaptive process.

Alternatively, a substitution may be contingent on or become entrenched by nonadaptive processes. Suppose that, instead of an environmental change, a mutation is fixed by drift at another site in the protein, changing the preference landscape at the site of interest. Such a shift in the landscape could increase the preference for alanine (an evolutionary Stokes shift) or decrease it (an evolutionary anti-Stokes shift). Given an increase in the preference for A, mutations away from A are unlikely to be fixed leading to its entrenchment. Alternatively, if the landscape shift resulted in a decrease in the preference of the resident amino acid such that another amino acid is

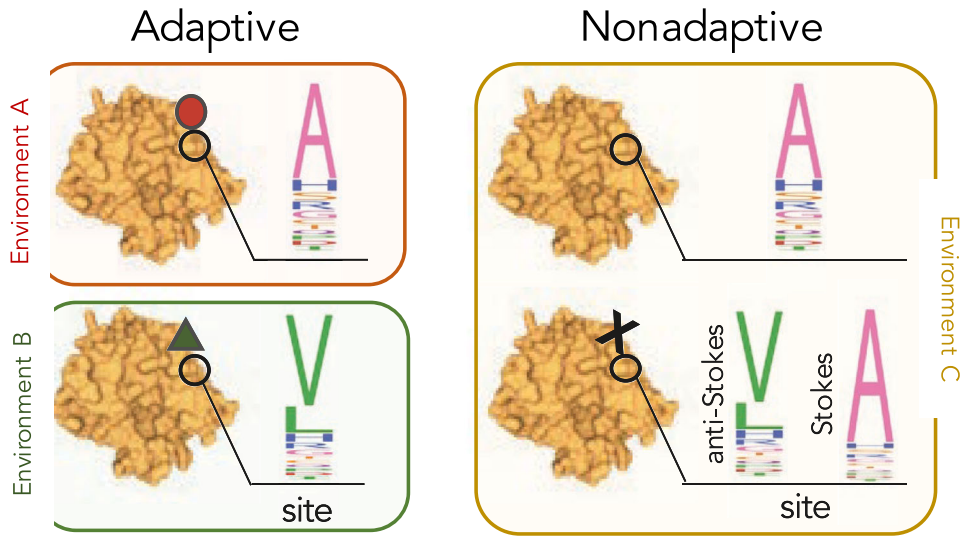


FIGURE 3 Shifts in amino acid preferences due to adaptive and nonadaptive processes. Suppose a change occurs in the protein's environment (e.g., a change in an interacting protein or a ligand; depicted by the red circle or green triangle), then the landscape shifts from having a strong preference for amino acid alanine (one-letter code A), to strongly preferring valine (one-letter code V). Nonadaptive evolution can also cause shifts in preferences. Following a substitution at another position in the protein (depicted with an X), the fitness landscape at a focal site could increase the preference for A or could change the ordering of amino acid preferences such that V is the most preferred residue. Evolutionary Stokes and anti-Stokes shifts are gradual phenomena that could in the long run lead to these example landscapes

the fittest at the site (e.g., V, Figure 3), then the subsequent fixation of a mutation encoding V is contingent on the change in the background sequence. These examples offer snap-shots of different preference landscapes. In natural protein evolution, these processes are dynamic and gradual over long periods.[3]

Shifted preferences can have significant consequences for protein evolution. Dobzhansky–Muller incompatibilities, where a mutation is neutral (or beneficial) in one protein but is pathogenic in a homologous protein, highlight the potential significance of shifted preferences on speciation.[19] Furthermore, entrenched substitutions play a significant role in maintaining molecular complexes.[20] It is, therefore, crucial to understand the drivers of shifts in preferences. The aim in this review is an attempt to quantify the magnitudes and frequencies of nonadaptive shifts in amino acid preferences.

3 | EVIDENCE OF PREFERENCE SHIFTS FROM MULTIPLE SEQUENCE ALIGNMENTS

A challenge with estimating shifts in preferences is that they are not directly observable in extant sequences. However, models which permit variation in site-specific preferences make explicit predictions that can be validated or refuted by patterns in natural alignments. Analysis of natural proteins often reveals evidence for temporal variation in replacement rates and homoplasy rates (reversions, convergence, and parallelism). Are these patterns explainable by nonadaptive processes, or are they the result of adaptive evolution? As reviewed below, in most instances, the observed patterns are consistent with predictions from nonadaptive epistatic models.

3.1 | Convergence rates

Convergence refers to the evolutionary phenomenon whereby similar traits emerge independently in multiple lineages. Convergence may occur at the phenotypic level, such as the origins of wings in bats and birds,[21] or echolocation in bats and toothed whales.[22] Phenotypic convergence is commonly viewed as evidence of adaptations of different lineages to similar environmental challenges.[23] Alternatively, molecular convergence, the emergence of identical states (nucleotide, codon, or amino acid) in two independent lineages, is not convincing evidence of adaptation since this could happen by chance owing to the limited number of permissable states at a site (four nucleotides, 61 codons, or 20 amino acids). Independent changes from the same ancestral state to the same derived state, are convergent substitutions that transpired in parallel (Figure 4).

Evidence of convergent substitutions abounds.[22,24–28] An adaptive explanation would suggest that convergent substitutions are due to similar selection pressures in different taxa. For example, Parker et al.[22] compared 22 mammalian genome sequences (encoding of 2,326 orthologous proteins) and reported a high number of convergent substitutions. They concluded that adaptive molecular convergence is widespread and explains the independent evolution of echolocation in bats and whales. However, their conclusions were challenged by two subsequent studies which reanalyzed their (and additional) data and found that convergence levels between bats and toothed whales are no greater than the levels of molecular convergence between bats and cows.[24,25] These studies highlight that rigorous assessments of the prevalence of adaptive convergence require properly formulated null models. Such null models allow us to assess whether it is necessary to invoke adaptive processes to explain observed patterns of substitution.

The simplest model for sequence evolution assumes equal substitution rates between states. This corresponds to the original Jukes and Cantor[29] model when describing substitutions between nucleotide states. When applied to amino acids, it is referred to as the Poisson model, assuming that all amino acids have the same fitness effect so that site-specific landscapes are uniform at all sites and are constant across time. The Poisson model predicts a relatively constant and low level of convergence rates as proteins diverge. However, evidence of convergence in natural datasets often exceeds the levels of convergences predicted by the Poisson model, and the level of convergence in natural alignments usually decreases as sequences diverge. Therefore, using the Poisson model as a null model, one might inaccurately reject the null in favor of an adaptive explanation. However, models which account for differences in rates of

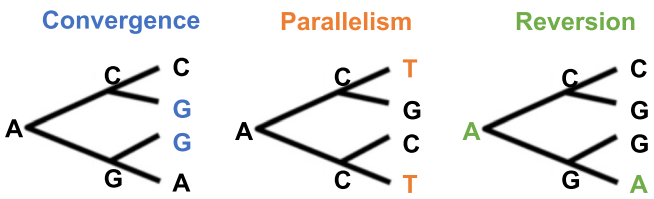


FIGURE 4 Examples of molecular homoplasy. Convergence refers to substitutions at independent lineages from different ancestral states to the same derived state. Parallelism refers to independent substitutions from the same ancestral state to the same derived state. Reversion refers to a change from a derived state back to an ancestral state

exchange among amino acids (for example, WAG[30]) and models that allow for variability across sites (for example, MutSel[31]) predict higher levels of convergence than the Poisson model, and declining levels with time. Nonetheless, rates of convergence inferred from natural alignments exceed the levels predicted by these heterogeneous models. A limitation of these models is that they do not account for epistasis. Various studies have independently shown that accounting for epistatic interactions leads to patterns and levels of convergence rates in line with observations in natural data.[26,27,32] In particular, their results highlight that understanding substitution patterns under epistatic models are imperative for accurately detecting adaptive evolution. In Box 1, we review two datasets with declining convergence rates. In both datasets, the observed patterns are consistent with nonadaptive epistatic dynamics.

Why do convergence levels decrease over time under epistatic models? To illustrate this, let us again consider Maynard Smith's word game analogy. The first position fitness landscapes are more similar when the background sequences have fewer differences (e.g., consider the first position landscapes given background sequences -ORD and -ORE; Figure 2). As more differences accumulate (e.g., -ORD and -ENE), the first position landscapes become more dissimilar. Similarly, in protein evolution, as sequences diverge the amino acid preference landscapes accumulate more differences.[32] As such, amino acid states that provide high fitness in one homolog may be unfit in a different background sequence. Nevertheless, structural or functional constraints could further limit variability in amino acid preferences across diverged proteins. The extent to which such restrictions limit variability in preferences, however, is still unknown.

BOX 1 Convergence rate: evidence of adaptations or expected under nonadaptive evolution?

Dataset: Thirteen orthologous mitochondrial proteins from 629 vertebrate mitochondrial genomes

- Goldstein et al observed declining levels of convergence rates with time in vertebrate mitochondrial proteins.[26]
- To dissect if the levels of convergence are evidence of adaptive evolution or are explainable by nonadaptive convergences, they simulated data under two substitution models: the WAG which is site- and time-homogenous but allows for difference in rates of exchange across sites; and a stability-mediated epistatic model which accounts for differences among sites and across time.[3,30]
- They found that the levels of convergence rates in the mitochondrial proteins were highly compatible with the levels expected under a nonadaptive epistatic model.

Dataset: 5,935 orthologous proteins from 12 fruit fly species

- Zou and Zhang report a large amount of variability in convergence rates across the different pairs of orthologous proteins. Convergence rates were higher in recently diverged proteins and declined with evolutionary distance.[27]
- To determine if the convergence levels are due to adaptive or nonadaptive process, they developed various evolutionary models and compared the expected rates to those observed in the natural proteins.
- The simplest model estimates gene-wide equilibrium amino acid frequencies, which are constant across sites and time. Based on this model, the observed number of convergences were significantly higher than the null expectation.
- They developed two additional substitution models both of which account for variation across sites by either grouping sites into classes with similar amino acid frequencies, or by assigning site-specific equilibrium frequencies. Under both these site-heterogeneous models, the observed convergence rates were significantly lower than predicted.
- Last, using simulations they showed that the lower rates of convergence in the empirical data compared to the site-heterogeneous null models is likely due to epistatic interactions.
- In conclusion, they found that the observed amounts of convergence is explainable by nonadaptive models which account for site- and time-heterogeneous process.

3.2 | Reversion rates

Reversion describes a return to an ancestral state during evolution (Figure 4). Molecular reversions are common in natural sequences.[33–35] More than a century ago, Muller[36,37] hypothesized that epistasis causes reversion rates to decrease with time. McCandlish et al.[38] proved that involvement in at least one epistatic interaction is sufficient to cause decreases in reversion rates, and that, in the absence of epistasis, reversion rates are constant through time.

Naumenko et al.[35] analyzed two datasets of genome-wide alignments from vertebrates (7,967 genes from 9 species) and insects (8,477 genes from 8 species). In both datasets, they observed decreases in reversion rates as sequences diverged, consistent with expectations under epistatic models.[38] Epistasis can lead to diminishing rates of reversion through (1) a nonadaptive increase in fitness for the derived residue (i.e., an evolutionary Stokes shift), or (2) a nonadaptive decrease in the fitness of the replaced residue.[35] Naumenko et al.[35] argued that the second effect is stronger and that "negative epistatic interaction with currently absent amino acids" is responsible for most of the observed declines in reversion rates.

3.3 | Replacement rates

Another signal commonly observed in natural alignments is changes in replacement rates over time, or *heterotachy*. Various adaptive and nonadaptive mechanisms can produce this signal. For example, evolution on a static site-specific fitness landscape, in the absence of both epistatic and adaptive processes, can lead to heterotachy.[6] On a static landscape, a chance fixation to a suboptimal amino acid is followed by a period of positive selection restoring the site to its optimal state, a process referred to as *nonadaptive shifting balance*.[6] Alternatively, heterotachy can also be caused by changes in site-specific fitness landscapes because of epistasis. Changes at other positions can lead to a more uniform fitness landscape having higher substitution rates, or a more rugged landscape with fewer opportunities for change.[39] Further, heterotachy may also occur because of changes in the S2S landscape resulting from an adaptive episode—the shift in the S2S landscape is often followed by a period of high substitution rates as the protein adapts to the new conditions.[4,6] Given the diversity of processes that can lead to heterotachy, accurate inference of the mechanisms at play in natural sequences is challenging.

Can heterotachy resulting from adaptive versus nonadaptive evolution be distinguished? Two studies have recently suggested that nonadaptive and adaptive processes cause idiosyncratic variations in replacement rates.[40,41] They hypothesized that epistasis causes a reduction in replacement rate with time, while adaptive evolution leads to increases in rates. The reason, they suggest, is that adaptive shifts in preferences often render the current state suboptimal for the new conditions. Positive selection will restore equilibrium through the subsequent fixations of beneficial nonsynonymous mutations, leading to an increase in substitution rate following the landscape shift. In contrast, nonadaptive evolutionary Stokes shifts increase the favorability of the resident amino acid. Such an increase in favorability leads to declining rates of replacement. However, the existence of an evolutionary anti-Stokes shift—where decreases in resident amino acid favorability lead to increases in replacement rates—challenges this claim.[2]

In this way, both adaptive and nonadaptive processes may lead to an increase in replacement rates over time. Nevertheless, we hypothesize that heterotachy caused by adaptive and nonadaptive processes can be differentiated. In the absence of adaptations, a balance is expected in the frequency and magnitude of both evolutionary Stokes and anti-Stokes shifts.[2] This balance suggests that under nonadaptive evolution, the proportion of sites that experience increases in replacement rates should be approximately equal to the proportion experiencing a decrease in rate. Alternatively, adaptive shifts will lead to an excess of sites with increased rates compared to the proportion of sites for which replacement rates decreased. This is akin to the expectations of the proportions of beneficial and deleterious substitutions under adaptive and nonadaptive processes. Under nonadaptive evolution, a balance exists in the proportions of beneficial and deleterious substitutions. However, following an adaptive change, the proportion of beneficial substitutions exceeds that of deleterious substitutions.[4,6] While the dynamics of landscape shifts under adaptive evolution are yet to be thoroughly investigated, we suspect that adaptive episodes will analogously lead to an excess in the proportion of sites undergoing increases in substitution rates relative to the proportion of rate-decreasing sites.

We summarize the results from three recent studies investigating changes in replacement rates in Table 1. In the reported datasets, the number of rate accelerating or decelerating sites is comparable—except for the hemagglutinin H3 subtype protein analysis in which a higher number of accelerating sites was observed (12 rate accelerating sites and only four decelerating sites). The sites with the largest increase in replacement rates were experimentally shown to affect antigenic properties.[40] Therefore, the observed increase in rates in the H3 protein may

TABLE 1 Number of rate accelerating sites is often equal to the number of rate decelerating sites, inline with expectations from nonadaptive epistatic models

References	Dataset	Rate increases	Rate decreases	Total number of alleles
Popova et al.[40]	H1 proteins from 1,613 strains	0	2	83
	N1 proteins from 2,015 strains	0	0	82
	H3 proteins from 1,832 strains	12	4	117
	N2 proteins from 1,996 strains	8	5	93
Stolyarova et al.[41]	Five mitochondrial genes across 3,557 metazoan species	28	21	42,637
Gelbart and Stern[42]	Nine proteins across 126 HIV-1/SIV strains	134	137	5,902

be a true signal of adaptive evolution. Nevertheless, the similar numbers of accelerating and decelerating sites in all other proteins are in line with the expectations from nonadaptive epistatic models.[2] Note, however, that the results presented in Table 1 are from a relatively narrow range of proteins, making it difficult to draw general conclusions. Future work establishing the differences and similarities in variability of replacement rates due to adaptive versus nonadaptive processes is warranted.

4 | EXPERIMENTAL EVIDENCE OF SHIFTS IN PREFERENCES

While the patterns discussed above—decreases in homoplasy rates with divergence levels, and patterns of heterotachy—are consistent with temporally varying preferences, they could have arisen by nonepistatic mechanisms. For example, inaccurate tree inference could lead to diminishing rates of convergence,[28] or nonadaptive shifting balance could lead to the observed heterotachy.[6] A more direct approach for inferring preference shifts driven by epistasis is to compare mutational effects across background sequences. If variations in preferences due to epistasis are minor, then a mutation should have a similar phenotypic effect regardless of the background sequence. Alternatively, if preferences depend heavily on sequence-context, then mutational effects will vary across different background sequences. Until recently, experimental methods were restricted in the number of mutations they can introduce.[43] Most studies performed one of three types of pairwise amino acid replacements (Figure 5): (a) *Forward mutations* by replacing the residue in an ancestral protein with a derived state; (b) *Backward mutations* which introduce an ancestral state into an extant protein; and (c) *Exchange mutations* by replacing the resident amino acid in one protein with the resident residue in an orthologous protein.

4.1 | Effects on stability

Protein stability is a holistic property determined by all residues in the polypeptide, the three dimensional configuration, and the physicochemical environment. Nevertheless, under the same environmental conditions, a stabilizing mutation in one sequence may be destabilizing in another. To investigate the dependence of the stability effect of a mutation on the background protein sequence, Ashenberg et al.[44] introduced the same mutations into a series of diverged homologs of the influenza nucleoprotein (NP). Specifically, they separately introduced six mutations (I186V, V239M, L259S, A280V, H334N, and G384R) into four NP homologs (Brisbane/2007, Aichi/1968, California/2009, and bat/2009). The level of sequence divergence relative to the Brisbane/2007 sequence is 8% with Aichi/1968, 10% with California/2009, and 28% with bat/2009. They observed that stability effects of mutations were conserved across background sequences: only a single mutation induced a substantial shift in stability effects (A280V). The substitution from A → V at site 280 was stabilizing in the context of the Brisbane/2007, Aichi/1968, and California/2009 sequences, but was destabilizing in bat/2009 NP. Analysis of their data revealed that the standard deviation in mutational effects on melting temperature across background sequences was 0.86°C, on average. Furthermore, the stability effects of mutations in the context of different homologous proteins were significantly correlated. However, correlations decreased as sequence divergence increased: the correlation in stability effects of mutations between Brisbane/2007 and Aichi/1968 (8% sequence divergence) was .90, falling to .89 in California/2009 (10% sequence divergence), and .82 in bat/2009 (28% sequence divergence).

To assess how stability effects of mutations change over time, Risso et al.[45] performed forward and backward substitutions between extant and ancestral reconstructions of thioredoxin proteins. Specifically, they assayed stability effects in the context of the extant *Escherichia*

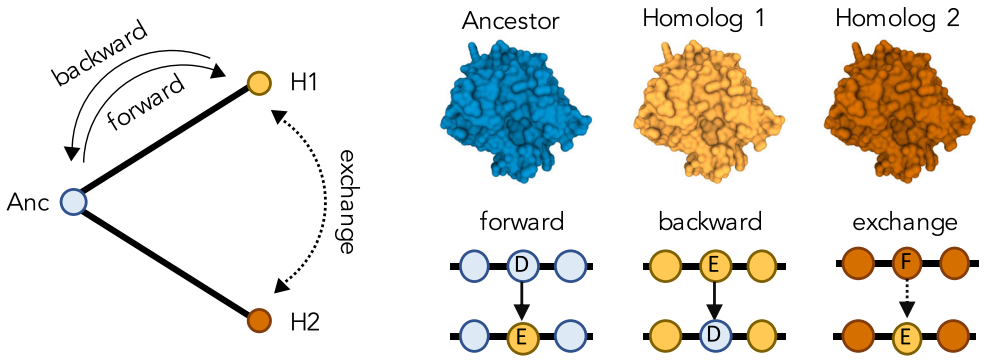


FIGURE 5 Diagram representing different mutation experiments. Forward substitutions place a derived amino acid into the context of an ancestral (Anc) sequence. Backward substitutions place an ancestral amino acid in the context of an extant sequence. Exchange substitutions refer to changing the resident amino acid in one homolog (e.g., E in H1) with the resident amino acid in another (e.g., F in H2). Forward and backward substitutions are shown in solid lines. Exchange substitutions are shown in dashed lines

coli protein and a "resurrected" protein present in the last bacterial common ancestor (LBCA). These proteins differ at 44% of sites. They introduced 21 mutations of the types E ↔ D, I ↔ V into both background sequences and assayed their effect on stability. The stability effects in the LBCA and *E. coli* thioredoxin proteins were strongly correlated (Pearson correlation coefficient of .89). Only 2 of the 21 mutations were stabilizing in one protein and destabilizing in the other. In general, stability effects of all mutations considered were within the range of ±1 kcal/mol. These results suggest that stability effects among biochemically similar amino acids (E and D, V and I) are conserved over long evolutionary time scales (approximately 4 billion years). To investigate the generalizability of this observation to biochemically dissimilar mutations, Risso et al.[45] introduced L ↔ K mutations across a series of ancestral thioredoxin proteins, and T ↔ M mutations across ancestral $\beta-$lactamases. Variability in stability effects was more pronounced in the L ↔ K and T ↔ M mutations than in the E ↔ D and V ↔ I mutations. Nevertheless, the most energetically preferred amino acid at a site remained the same in the extant and ancestral proteins.

The experimental studies reviewed above investigated the stability effects of a limited number of mutations. Alternatively, simulations of stability-constrained evolution allow for a more comprehensive assessment of stability effects across a wide range of background sequences.[3,13] Shah et al.[13] simulated the evolution of the lysine-arginine-ornithine-binding periplasmic protein (argT) using the force-field approach FoldX to estimate stability. They performed forward and backward mutations in silico and assayed the stability effects across all background sequences. They observed that variability in stability effects was common in frequency, yet minor in magnitudes. On average, stability effects were within 0.8 kcal/mol. In summary, theoretical and experimental investigations reveal that stability effects of mutations are conserved across background sequences, consistent with the expectation that fitness effects are often nearly neutral at mutation-selection-drift equilibrium.[46,47]

4.2 | Effects on function

The previous results suggest that stability effects of mutations are conserved across diverged sequences. Are functional effects of mutations similarly conserved, or is protein function highly attuned to the background sequence such that functional effects of mutations differ substantially across background sequences?

Lunzer et al.[48] were amongst the first to investigate the functional effects of mutations in orthologous proteins. They individually introduced 168 mutations into the wild-type *E. coli* isopropyl malate dehydrogenase (IMDH) protein and assayed their impact on enzyme performance (k_{cat}/K_m). At each site, they performed exchange mutations with the resident amino acids present in the *Pseudomonas aeruginosa* IMDH homolog. The vast majority of single mutant enzymes (104/168) performed similarly to the wild-type IMDH proteins, suggesting that functional effects of mutations are conserved.

Emlaw et al.[49] compared the effects of mutations on single-channel conductance using human muscle-type acetylcholine receptor (AChR) and an ancestral AChR (the AChR present in the last common ancestor between humans and cartilaginous fish). The proteins differed at 36% of sites. At two preselected sites where the resident amino acids differed between the two proteins (sites 2 and 6), they performed backward substitutions, placing the ancestral amino acids into the human AChR

(mutations G2T and F6S). They also performed forward substitutions, placing the derived amino acids into the ancestral sequence (mutations T2G and S6F). They also introduced the double mutants into both the extant and ancestral proteins. Analysis of their data revealed high concordance between the effects of the studied mutations in the different background sequences (Pearson correlation was .90).

Starr et al.[50] performed forward and backward replacements between a heat shock protein 90 (Hsp90) ATPase domain present in modern *Saccharomyces cerevisiae* (ScHsp90) and a reconstructed deep eukaryotic ancestor, ancAmoHsp90—the reconstructed Hsp90 sequence of the common ancestor of Amorphea, a eukaryotic supergroup comprised of animals, fungi, amoebae and other protists (see Reference 51 for taxonomic definition). In particular, their analysis focused on the N-terminal domain (NTD). The ancestral and extant NTDs differ at 60 of 221 sites (27% sequence divergence). They individually introduced each ancestral amino acid into the extant ScHsp90 protein and each derived state into ancAmoHsp90. Then, they estimated the fitness of yeast cells carrying the mutant proteins by measuring the change in the ratio of a mutant to wildtype frequency over time. In this way, they surveyed the effects of mutation on both function and stability. Approximately 48% of derived states reduced fitness when placed in the context of the ancestral NTD, 32% were neutral, and 20% were beneficial. When placed in the modern NTD, 92% of ancestral amino acids were deleterious, 7% were neutral, and 1% were beneficial. Across all mutations studied, 77% had different impacts on fitness depending on the background sequence. However, the effects of most mutations were minor: the average selection coefficient was −.02, and −.01 for backward and forward substitution, respectively. (A selection coefficient of .01 means that the mutation will be effectively neutral if the effective population size is less than 100.) Note that the relatively small selection coefficient does not imply that epistasis plays a minor role in protein evolution. Even a relatively small degree of nonadditivity in the effects of mutations can have a considerable impact on evolutionary processes: epistasis has a strong influence on the accessibility and probability of evolutionary trajectories.[13,52,53]

4.3 | Quantifying the frequency and magnitude of shifts in preferences using deep mutational scanning

The previously discussed studies were limited to a small number of mutations. However, recent advancements, known collectively as deep mutational scanning (DMS), allow us to estimate the fitness effect of all single amino acid mutations at many (or all) sites in a protein.[43,54] First, a single-mutant library of proteins is created. The mutants are then subjected to a selection or screen in which the frequency of each genotype in the library is measured using deep sequencing. Fitness can then be estimated from the frequency measures. One approach is to evaluate a mutant's frequency relative to the wildtype over time as a measure of fitness.[50] Others have used the relative frequency of a mutant pre- and post-selection as a measure of the mutant's fitness.[5] More sophisticated Bayesian approaches, which correct for low sequencing depth have also been developed (see Reference 5 for a detailed description of models used to analyze DMS data and software implementations). While DMS approaches are a powerful tool for assessing the extent of shifts in amino acid preferences, the level of experimental noise is often high. Site-specific landscapes estimated from replicate experiments can have correlation coefficients as low as .59.[55,56]

Despite its recency and potential limitations, DMS methodologies have been used to estimate site-specific fitness landscapes in many proteins in various organisms. Livesey and Marsh,[57] report on the results from 31 publicly available DMS datasets: 13 from human proteins, 9 from bacterial proteins, 5 from yeast proteins, and 4 viral proteins. However, only four studies have applied DMS to homologous proteins.[55,58–60] Six datasets from these four studies are available to compare site-specific preferences across different background sequences (Table 2). Three studies were carried out in viruses.[55,58,59] The fourth study[60] compared site-specific fitness landscapes in orthologous indole-3-glycerol phosphate synthase (IGPS) proteins present in the archaeon *Sulfolobus solfataricus* (ssIGPS) and in two bacteria: *Thermotoga maritima* (TmIGPS) and *Thermus thermophilus* (TtIGPS). Collectively, the studies compare site-specific landscapes across sequences with as little as 6% and up to 73% sequence divergence.

There are broadly two ways of comparing site-specific landscapes across different sequences. The first approach is to calculate correlation coefficients between landscapes. This has been done in two ways (Figure 6): (a) calculate the landscape correlation at homologous sites, and report the mode of the correlation coefficient distribution (R_{mode}; Figure 6a); or (b) concatenate all landscapes and estimate a single overall correlation coefficient (R_{overall}; Figure 6b). Chan et al.[60] used the first approach and found that site-specific landscapes were significantly correlated (with modes ranging from .62 and .72; Table 2). Alternatively, Bloom and colleagues report the overall correlation from the second approach: R_{overall} ranged from .36 to .72.[55,58,59] It is currently unclear if

TABLE 2 Site-specific preference landscapes estimated across diverged background sequences are positively correlated

References	Organism	Protein	Comparison	Seq. length[a] (# sites[b])	% div	Correlation between	Correlation within	Prevalence
Doud et al.[58]	IAV	NP	H1N1–H3N2	497 (497)	6%	.78[c]	.83[c]	2.8% (FDR of 0.05)
Haddox et al.[55]	HIV	env	BF520–BG505	836 (659)	14%	.57–.58[d]	.59–.78[e]	4.6% (FDR of 0.01)
Lee et al.[59]	IAV	HA	H1N1–H3N2	566 (566)	58%	.36–.47[d]	.69–.82[e]	–
Chan et al.[60]	*S. Solfataricus* (Ss)	IGPS	SsIGPS–TtIGPS	271 (80)	65%	.72[f]	.94[f]	–
	T. thermophilus (Tt)		SsIGPS–TmIGPS	267 (80)	70%	.62[f]		
	T. maritima (Tm)		TmIGPS–TtIGPS	277 (80)	73%	.62[f]		

Note: Listed are the Pearson correlations between landscapes within replicate experiments (correlations within), and correlations between landscapes estimated in different background sequences (correlations between). Prevalence is estimated from the $RMSD_{corrected}$ approach.

[a]Pairwise alignable sites.

[b]Number of mutated sites.

[c]$R_{overall}$ between replicate-averaged site-specific landscapes. Within replicate correlations are based on comparison with site-specific landscape estimates from a previous study.[61]

[d]Range of $R_{overall}$ overall replicate pairs between homologs.

[e]Range of $R_{overall}$ overall replicate pairs within homologs.

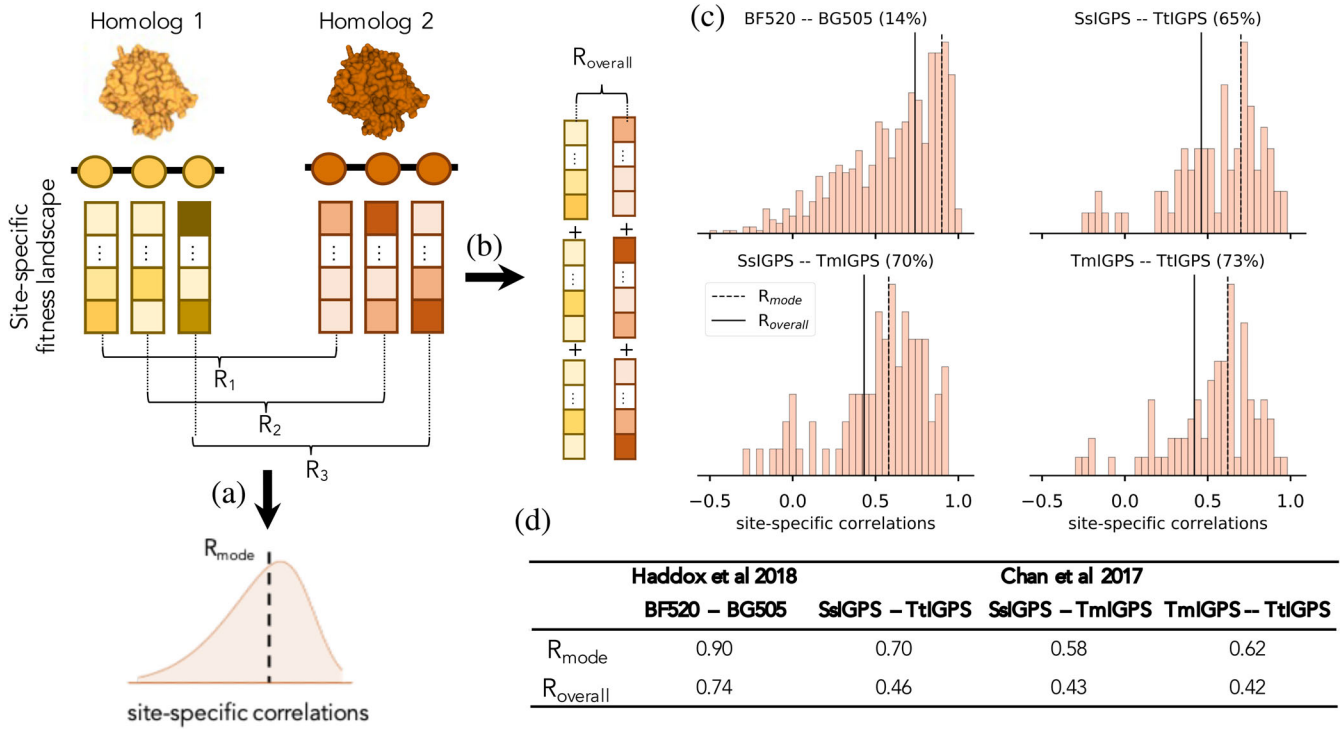
[f]R_{mode}.

both approaches lead to similar correlation estimates and hence similar biological conclusions.

To compare the two correlation approaches, R_{mode} and $R_{overall}$, we reanalyzed the datasets from Chan et al.[60] and Haddox et al.[55] using both methods. Note that, Haddox et al.[55] conducted three replicate experiments for each homologous protein (BF520 and BG505). It is valuable to obtain the across-replicate average landscapes prior to obtaining correlations (see Box 2 for more details). We report the correlations between site-specific landscapes given the different background sequences in Figure 6c,d. It is clear from this analysis that R_{mode} and $R_{overall}$ can differ; specifically, $R_{mode} > R_{overall}$ in the four datasets. The largest difference is observed in the TmIGPS—TtIGPS comparison where R_{mode} and $R_{overall}$ differ by 0.20. Because fitness profiles can be expected to vary substantially over sites, conditioning on a site by reporting site-specific correlations may be more statistically robust and is more informative regarding the dynamics at a site. For example, it is evident from the site-specific correlation distributions that most landscapes correlate strongly ($R > .5$). However, some sites have landscapes that are negatively correlated. A negative correlation of the preference landscape given different genetic backgrounds indicates substantially shifted amino acid preferences and suggests differing functional or structural constraints in the respective proteins. For these reasons, considering outliers in the entire distribution of site-wise correlations may be preferable for inferring specific locations where preferences have shifted substantially.

In order to accurately detect shifts in preferences using DMS data we must account for high amounts of experimental noise. Therefore, a second approach for quantifying shifts in amino acid preferences compares the distance between two landscapes using the Jensen-Shannon distance metric[58] (see Box 3 for detailed discussion). Briefly, the distance is equal to zero when amino acid preferences are identical, and is one if the preferences are dissimilar. The distance approach accounts for the level of variability in site-specific landscapes due to experimental noise by estimating the average root-mean-square distance within replicate experiments ($RMSD_{within}$). Similarly, the distance between site-specific landscapes in homologs is calculated ($RMSD_{between}$). The magnitude of shift at a site ($RMSD_{corrected}$) is then calculated as the difference between $RMSD_{between}$ and $RMSD_{within}$. In summary, the $RMSD_{corrected}$ value at each site provides a measure of the magnitude of the shift in preference while calibrating for experimental noise.

The $RMSD_{corrected}$ approach can be used to quantify the prevalence of significantly shifted sites. To do this, a null distribution of $RMSD_{corrected}$ values is generated through an exact permutation test by reassigning site-



	Haddox et al 2018	Chan et al 2017		
	BF520 – BG505	SsIGPS – TtIGPS	SsIGPS – TmIGPS	TmIGPS – TtIGPS
R_{mode}	0.90	0.70	0.58	0.62
$R_{overall}$	0.74	0.46	0.43	0.42

FIGURE 6 Different approaches for comparing correlations between site-specific landscapes across different background sequences. The first approach (a) estimates the correlation between landscapes at homologous sites given different background sequences and reports the mode of distribution (R_{mode}). The second approach (b) concatenates all site-specific landscapes and estimates an overall correlation value ($R_{overall}$). (c) Distribution of site-specific correlation values from four DMS experiments. The BF520–BG505 dataset is from Haddox et al.[55] The remaining datasets are from Chan et al.[60] Percentages in parentheses are the percent sequence divergence between the two proteins. (d) Reports the values of R_{mode} and $R_{overall}$ from the four datasets. Reported correlations for the BF520–BG505 dataset are from site-specific fitness landscapes averaged over replicate experiments

specific landscapes among the two protein groups. If preferences have not shifted significantly between the two proteins, then the true distribution of $RMSD_{corrected}$ values should be similar to the null distribution. This method can be used to identify sites for which the null hypothesis of no shifts is rejected. Note that, permutation tests can be conservative because they construct a null distribution from data that may instead support the alternative hypothesis. As such, this approach may be susceptible to high false negative rates.

Doud et al.[58] performed DMS on two homologs of influenza A virus (IAV) NPs in the H1N1 and H3N2 viral strains. The proteins differed at 6% of sites. Using the $RMSD_{corrected}$ approach, they found that only a modest fraction of sites exhibited significant shifts in amino acid preferences: at a false discovery rate of 0.05, 14 of 497 sites (2.8%) showed evidence of significantly shifted preferences. Haddox et al.[55] used the same method to quantify the magnitude and prevalence of shifted preferences between homologous HIV envelope (env) proteins that differ at approximately 14% of sites. Only 30 of the 659 sites (4.6%) showed evidence of significantly shifted preferences (at an FDR of 0.01). Lee et al.[59] performed a similar analysis between homologous hemagglutinin (HA) proteins present in influenza viruses H1N1 and H3N2. The proteins were highly diverged, having 58% sequence divergence. The number of significantly shifted sites was not reported. However, it is evident from the distribution of $RMSD_{corrected}$ (Figure 7c in Lee et al.[59]) that a large number of sites had significantly shifted preferences. Also, the magnitude of the shifts was more pronounced than in other DMS studies. For example, the largest $RMSD_{corrected}$ reported in Doud et al.[58] was 0.45, whereas $RMSD_{corrected}$ values were as high as $\approx$ 0.8 between the hemagglutinin homologs.

A challenge with assessing shifts in preferences using the correlation approaches is that, while it is clear that correlations between landscapes inferred from homologs are lower than correlations from biological replicates, it is unclear if the observed decreases are statistically significant. As such, the $RMSD_{corrected}$ approach has been informative for inferring significantly shifted preferences.[55,58] Nonetheless, a limitation of the $RMSD_{corrected}$ approach is that it cannot distinguish between instances where the order of amino acid preferences has changed versus cases where there is an intensification (or relaxation) of

BOX 2 Higher correlations when averaging over replicate experiments

Deep mutational scans can display high levels of experimental noise. Therefore, triplicate experiments are usually conducted for a given protein. Let P and Q be the true site-specific (or concatenated) fitness landscapes given different background sequences. Then let P^r and Q^r be the fitness landscapes estimated from a DMS experiments r and s such that $P^r = P + e^r$ and $Q^s = Q + d^s$, where e^r and d^s are the measurement errors. If these are uncorrelated then,

$$\mathrm{Cov}(P^r, Q^s) = \mathrm{Cov}(P, Q)$$

but, $\mathrm{Var}(P^r) = \mathrm{Var}(P) + \mathrm{Var}(e^r)$ and $\mathrm{Var}(Q^s) = \mathrm{Var}(Q) + \mathrm{Var}(d^s)$. Thus, the correlation for a replicate pair (r, s) is

$$\begin{aligned}\mathrm{Corr}(P^r, Q^s) &= \frac{\mathrm{Cov}(P^r, Q^s)}{\sqrt{\mathrm{Var}(P) + \mathrm{Var}(e^r)} * \sqrt{\mathrm{Var}(Q) + \mathrm{Var}(d^s)}} \\ &= \frac{\mathrm{Cov}(P, Q)}{\sqrt{\mathrm{Var}(P) + \mathrm{Var}(e^r)} * \sqrt{\mathrm{Var}(Q) + \mathrm{Var}(d^s)}} \\ &\leq \frac{\mathrm{Cov}(P, Q)}{\sqrt{\mathrm{Var}(P)} * \sqrt{\mathrm{Var}(Q)}} \\ &= \mathrm{Corr}(P, Q)\end{aligned}$$

This shows that correlations estimate from replicate pairs of experiments will be less than or equal to the true correlation. The above argument also holds for across-replicate averaged landscapes $\overline{P}$ and $\overline{Q}$.

$$\mathrm{Var}\left(\overline{P}\right) = \mathrm{Var}(P) + \mathrm{Var}(e^r)/3$$

$$\mathrm{Var}\left(\overline{Q}\right) = \mathrm{Var}(Q) + \mathrm{Var}(d^s)/3$$

However, the denominator term, causing the underestimation, is smaller for averaged landscapes. For example, consider the Haddox et al.[55] study, where they performed DMS triplicate experiments for envelope proteins present in HIV stains BF520 and BG505. The correlations between replicate experiments, $\mathrm{Cov}(P^r, Q^s)$ were less than .58. However, the correlation between across-replicate average landscapes was .74.

In summary, it is valuable to average prior to obtaining correlations. If errors in approximating the landscapes are uncorrelated, the covariance does not change by averaging but the variance contributions due to errors in approximation are reduced giving a better approximation to the correlation of interest that one would obtain had there been no variation over replicates.

selection between sequences. An example of this is provided in Figure 7. Amino acid alanine (one-letter code A) is the most preferred residue at site 512 in both homologs of the env protein.[55] However, site 512 is more mutationally tolerant in the context of the BG505 sequence versus the BF520 background. Conversely, at site 288, there is a clear shift in the ordering of amino acids. Despite having different shifted dynamics, the $\mathrm{RMSD}_{\mathrm{corrected}}$ approach estimates a similar degree of shift at sites 288 and 512. Alternatively, the Pearson correlation between landscapes is substantially lower for site 288 (Figure 7), highlighting that the correlation approach might be more suitable for identifying sites having different preferred amino acids given different background sequences.

DMS is a promising tool for quantifying the magnitude and prevalence of shifted amino acid preferences. In addition to the analyses discussed above, data from DMS can be used to assess multiple additional questions: How often is a substitution deleterious in one protein but beneficial in another? How often does the most preferred amino acid at a site differ across background sequences? How often are the detected shifts due to a reordering of the preferred amino acid versus a relaxation (or intensification) of selection pressure? Answers

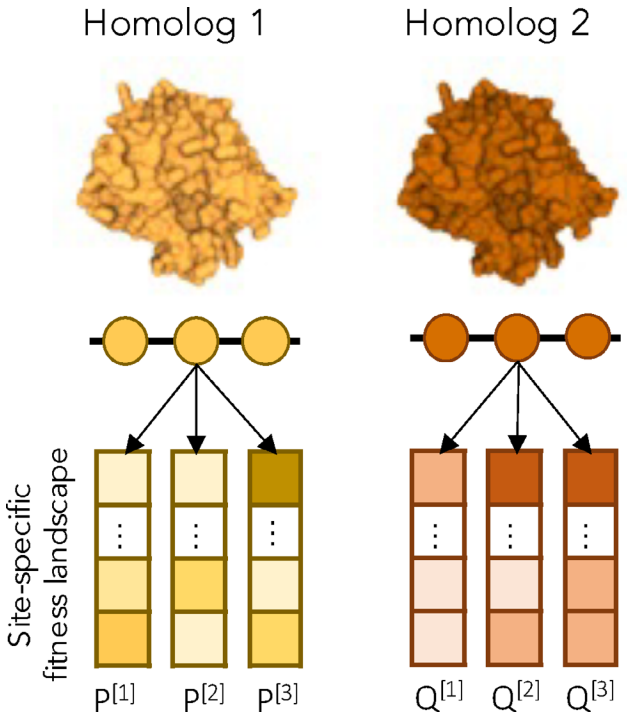
BOX 3 Quantifying shifts in DMS datasets while controlling for experimental noise

DMS methodologies offer a powerful tool for assessing the effect of mutations. However, the level of experimental noise may be problematic; correlations between identical replicates can be as low as .59, matching correlation coefficients observed across different background sequences.[54,55] Therefore, quantifying the extent and prevalence of shifts in preferences must be calibrated to the observed level s of experimental noise.

Doud et al.[58] used the Jensen–Shannon divergence (JSD) to quantify the level of similarity (or dissimilarity) between the fitness landscapes at homologous sites given different background sequences.[58] Let P and Q be the site-specific fitness landscapes at a site given the background sequences H1 and H2, respectively. Then,

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathbf{D}(P||A) + \frac{1}{2}\mathbf{D}(Q||A)$$

where $A = \frac{1}{2}(P+Q)$ is the average fitness landscape and $\mathrm{D}(P \,||\, A) = \sum_i \mathrm{P}_i \log(P_i/Q_i)$ is the Kullback–Leibler divergence. Let $d(P,Q) = \sqrt{JSD(P|Q|)}$, such that $d(P,Q)$ is a metric of the distance between landscapes P and Q. The utility of $d(P,Q)$ is that it is symmetric and ranges from 0 (identical distributions) to 1 (dissimilar distributions).



Replicate experiments yield different landscape estimates. To quantify the level of variability within replicates, calculate the average root-mean-squared distance at a site within replicate experiments:

$$\mathrm{RMSD}_{\mathrm{within}} = \frac{1}{2}\sqrt{\frac{1}{n_P}\sum\nolimits_{r,s \in N_P} d(P^r,P^s)^2} + \frac{1}{2}\sqrt{\frac{1}{n_Q}\sum\nolimits_{r,s \in N_Q} d(Q^r,Q^s)^2}$$

where P^r is the estimated landscape at a site in replicate r, N_P is the set of nonredundant pairwise comparisons within replicates (e.g., given three replicate experiments, $N_P = \{(1,2), (1,3), (2,3)\}$), n_P is the number of comparisons, and the respective definitions for Q^r, N_Q, and n_Q. Then, calculate the root-mean-square distance between landscapes in different background sequences

$$\mathrm{RMSD}_{\mathrm{between}} = \sqrt{\frac{1}{n_{P,Q}} \sum\nolimits_{r,s \in N_{P,Q}} d(P^r, Q^s)^2}$$

where $N_{P,Q}$ is the set of nonredundant pairwise comparisons between replicates (e.g., given three replicate experiments for each background sequence, $N_{P,Q} = \{(1,1), (1,2), (1,3), (2,1) \ldots, (3,3)\}$), and $n_{P,Q}$ is the number of comparisons. The magnitude of preference change after correcting for site-specific noise is calculated as

$$\mathrm{RMSD}_{\mathrm{corrected}} = \mathrm{RMSD}_{\mathrm{between}} - \mathrm{RMSD}_{\mathrm{within}}$$

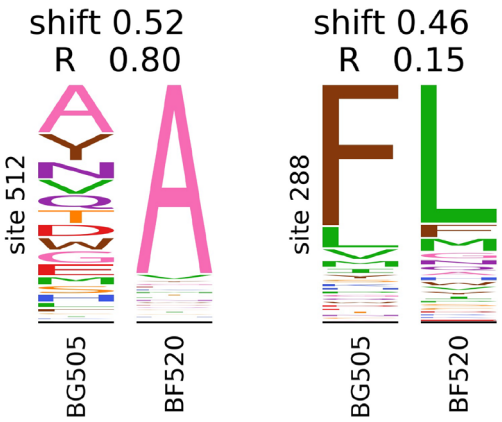


FIGURE 7 Correlation approach is better at identifying a reordering of amino acid preferences compared to the $\mathrm{RMSD}_{\mathrm{corrected}}$ approach. Site-specific preference landscapes in homologous envelope proteins in HIV strains BG505 and BF520. Shown are the across-replicate average preference landscapes at a site. The reported shift is the $\mathrm{RMSD}_{\mathrm{corrected}}$ values. The reported R value is the Pearson correlation coefficients between site-specific preference landscapes. Data obtained from Haddox et al.[55] under the Creative Common Attribution license

to these questions can be illuminated using data from DMS.

5 | LIMITATIONS

The studies reviewed above suggest that temporal variability in amino acid preferences is usually minor in magnitude and low in frequency. However, each of the methods used for inferring preference shifts has potential limitations. Detecting variations in rates of homoplasy or replacement rates in natural alignments are indirect ways of assessing preference shifts. While theoretical models suggest that epistasis could result in the observed signal, other mechanisms may also be at play.[28] Alternatively, DMS approaches allow for a more direct assessment of site-specific preferences in different background sequences. These approaches offer snap-shots of preference landscapes in the context of different sequences but tell us little about the trend of change over time. For example, we cannot use current DMS data to assess if changes in preference are abrupt or gradual. Nevertheless, comparing preference landscapes between ancestral and extant proteins (as done in Starr et al.[50]) to track how preferences change over time is valuable for understanding trends in preference shifts.

Ancestral reconstructions of ancient proteins are widely used in phylogenetics and are of particular importance for assessing the degree of change in site-specific preference landscapes over time.[45,49,50] Nevertheless, the models used to infer ancestral character states often assume stationary amino acid frequencies over time and across sites, and no changes in S2S landscapes (i.e., no adaptive change) which may lead to biases in the ancestral reconstructions.[62,63] Thus reconstruction could have some degrees of systematic error or bias that has an asymmetric effect on the background sequences that are evaluated. If, say, 30% of states in an ancestral protein must be reconstructed, then biases could affect up to 30% of the ancestral sequence but 0% of the extant sequence within the various experiment settings. So, when comparing the effects of changes between ancestral and extant proteins there could be an asymmetric impact of such biases. The potential impact on interpretation is further complicated since reconstruction biases tend to overestimate thermostability of ancestral proteins.[62] While the narrow sense interpretation of the experiments will be correct, they might not as easily be generalized to the actual ancestors such as the LBCA. Investigating the mutational effects across a sample of ancestral sequences with high posterior probabilities could potentially reveal disparity in preference landscape inference, leading to different conclusions regarding the magnitudes and shifts in amino acid preferences over time. Nevertheless, results from comparative analyses of amino acid preferences in extant, homologous proteins are not subject to the same biases as ancestral reconstructions.[55,58–60] The conclusions from such

analyses reveal that preferences are often conserved in homologous proteins, inline with conclusions based on reconstructed proteins.

Currently, there is limited availability of data from DMS experiments that can be used to assess shifts in site-specific preferences. There are only four studies that compared preference landscapes between homologous protein sequences (Table 2), and only one of these compares orthologous bacterial and archaeal proteins.[60] The remaining three studies were conducted in viruses, specifically RNA viruses. The high mutation rates in RNA viruses may have selected for loosely packed protein structures, which buffer the deleterious effects of mutations.[64] This would suggest that the low levels of mutational effects observed in these experimental settings may not generalize to nonviral proteins. This has led to concerns regarding the utility of viral DMS data in more generally assessing levels of preference shifts.[65] However, results from Chan et al.[60] corroborate that drastic shifts in preference landscapes are usually rare in nonviral proteins even at high levels of sequence divergences. Furthermore, Ferrada[66] curated a dataset of 124 pairs of homologous proteins (sequence divergences ranged from 0 to 100%) and computationally estimated site-specific landscapes using FoldX. Using the $RMSD_{corrected}$ approach, they observed that the number of sites with significantly shifted preferences increases with sequence divergence. Nevertheless, even at 100% sequence divergence on average less than 30% of sites had significantly shifted preferences. This study only modeled the effects of stability. Natural proteins are affected by additional structural constraints, beyond just stability, that influence preferences in ways that are only marginally dependent on background sequence. Therefore, the percentage of sites with substantial shifts in preference in natural proteins are likely to be even less than 30%.

6 | CONSEQUENCES OF SHIFTS FOR TIME-HOMOGENOUS EVOLUTIONARY MODELS

One way of deducing information about evolutionary processes is to analyze multiple sequence alignments with a quantitative model of sequence evolution. Two widely used classes of evolutionary models are phylogenetic models used to infer relationships between taxa and ω models used to estimate selection intensity. Inference procedures for either class of models often assume that the evolutionary process is identical across sites and constant through time. Specifically, most models assume (a) independent evolution across sites, (b) time-homogeneous substitution processes, and (c) a common vector of stationary frequencies; assumptions that are all violated in the presence of epistasis.

Various amendments have been applied to allow for heterogeneity (spatial and temporal) in the evolutionary process in both phylogenetic and ω models. However, due to the difficulty in tractably modeling co-dependencies among sites, models are limited in the extent of heterogeneity that can be accommodated. In practice, inference procedures model among-site heterogeneity through a mixture model with different substitution processes as classes in the mixture, and can allow for temporal changes in the substitution process at prespecified branches along the tree,[10,11,67] or using a covarion-like process.[68–70] There has also been significant development on site-heterogeneous models informed by protein structure.[71–75] More recently there has been a push toward using experimentally informed evolutionary models where site-specific substitution processes are informed by data from DMS.[61,76,77] While these models offer improved likelihood scores over more traditional approaches, they are limited in applicability to the currently small number of proteins for which DMS data is available.

While the challenges associated with allowing for temporal and spatial heterogeneity place a high barrier for their widespread incorporation into inference procedures, it is nonetheless of paramount importance to understand how they may bias our inferences. To this end, recent studies have advocated for the use of models of protein evolution with plausible levels of spatial and temporal heterogeneity as a tool for assessing the accuracy of inference in the face of realistic levels of heterogeneity.[6,8,70,78,79] Simulations of stability-informed models recapitulate levels of both spatial and temporal heterogeneity present in real data.[8] They are therefore a powerful tool for assessing inference accuracy. To this end, sequences are first generated under a stability-constrained evolutionary model. The simulated sequences are then analyzed using traditional inference procedures. The true parameter values, predicted from the generating model, are then compared to the inferred parameters to assess inference accuracy.

Using the procedure outlined above, it is evident that traditional ω models underestimated levels of among-site heterogeneity; ω models estimated only 2–4 rate classes when a much richer distribution of rate classes (>100) is present in the true generating process.[8] Nevertheless, the inferred rates corresponded to the most common substitution rates across sites. Furthermore, inclusion of a covarion-like component in the substitution model, allowing rates at sites to vary over time, fit the data significantly better. These results suggest that ω models need not explicitly include epistatic interactions for reasonable inference of selection pressure when averaging over time

and sites, and that allowing for a covarion-like component seems to capture temporal heterogeneity in rates arising due to epistasis.[8]

The procedure outlined above has not yet been implemented to assess the sensitivity of phylogenetic inference to extensive and persistent levels of heterogeneity due to nonadaptive stability-constrained epistasis. However, the literature assessing the accuracy of phylogenetic inference in the face of temporal and spatial heterogeneity "by-any-means" is vast. Simulations show that ignoring temporal heterogeneity can induce systematic errors in phylogenetic inference, including topological and branch length inaccuracies.[80–84] However, it remains unclear if the level of heterogeneity arising from nonadaptive epistatic processes is substantial enough to similarly bias our phylogenetic inferences. Assessing the implications of epistasis on phylogenetic inference is a fruitful avenue for future research.

In contrast with the relatively minor changes in preferences over time, differences in amino acid preferences among sites is substantial.[7] Models that accommodate among-site heterogeneity fit the data significantly better than site-homogeneous models.[77] This leads to the question: How can site-specific fitness profiles be estimated? There are currently two approaches for obtaining site-specific fitness landscapes: (a) they can be statistically inferred from large multiple sequence alignments (e.g., Rodrigue and Lartillot[16]), or (b) experimentally obtained from deep mutational scans (e.g., Hilton and Bloom[77]).

A new approach, informed by developments in the field of systems biology, might be worth exploring. Various computational variant effect predictors (VEPs) have recently been developed to predict the effects of mutations in a given protein sequence, often for clinical applications. In a recent study, Livesey and Marsh[57] compared the performance of 46 different computational VEPs to data obtained from DMS. These VEPs rely on various structural, evolutionary, and biophysical features (see Reference 57 for details of the different VEPs). The best performing VEP was DeepSequence,[85] an unsupervised machine learning approach. DeepSequence had an average correlation coefficient between predicted and observed (DMS) landscapes equal to .43 across all human proteins and .46 across all nonhuman proteins. While these correlation coefficients are low, it is relevant to note that the average Pearson correlation between different DMS studies on the same protein is only .66,[57] and correlations between replicate experiments can be as low as .59.[55,56] A noteworthy prerequisite of the DeepSequence method is that it necessitates the availability of large multiple sequence alignments. For proteins where a large alignment is not available, other VEPs that rely on structural or biophysical features, such as DEOGEN2[86] and SNAP2,[87] may be preferable. As with any supervised approach, DEOGEN2 and SNAP2 have potential limitations related to overfitting of the training dataset. Nevertheless, both methods performed well against diverse, independently curated DMS datasets—from viral, eukaryotic, and bacterial proteins—highlighting their potential utility to generally estimate mutational effects.

Site-specific fitness landscapes can be estimated from VEPs and used to inform evolutionary models. For example, site-specific frequency landscapes can be estimated from the site-specific fitness landscapes and provided to phylogenetic models, similar to the phylogenetic application of DMS data.[76,77] Alternatively, fitness values can be used directly in models of sequence evolution to specify the rates of substitutions between codons or amino acids. Bloom[61] proposed two heuristic approaches of converting site-specific fitness landscapes to fixation probabilities. These approaches were first developed in the context of DMS data but can be used to estimate fixation probabilities from landscapes predicted from VEPs.

While we do not yet have a complete understanding of the degree of temporal shifts in most proteins, the reviewed studies suggest that they are usually minor in magnitude at most sites and that only a small fraction of sites have significantly shifted preferences. These consistent yet minor perturbations in preferences have significant consequences for sequence evolution[13,52,53] and can lead to variation in rates across time.[2] However, most inference models assume constant preferences. Evidence is emerging highlighting the value of accounting for temporal heterogeneity in inference procedures using a covarion-like process (e.g., Jones et al.,[70] Lu and Guindon[88]). Therefore, allowing for temporal variability (using a covarion-like component) in addition to allowing preferences to vary across sites (estimated experimentally or computationally) might lead to better models of protein evolution.

7 | CONCLUSIONS

From the foregoing, it is clear that nonadaptive processes can alter site-specific amino acid preferences. Experimental studies suggest that at high sequence divergence levels only a small proportion of sites experience significantly shifted preferences, while at most sites there is only a small quantitative perturbation to the observed amino acid preference.[55,58–60] Extensive computational studies corroborate this conclusion.[13,66] Furthermore, pairwise amino acid exchange mutations between highly divergent sequences often have only minor differential effects on fitness,[50] function,[48,49] and protein stability.[44,45] Together these results suggest that amino acid preferences at most

sites vary slightly but are usually conserved over long evolutionary time scales. Nevertheless, the frequent, but small, changes in amino acid preferences leave an identifiable footprint in natural sequences: decreases in convergence rates,[26,27] reversion rates,[35,38] and variation in replacement rates[40–42] with time, and can have significant implications for protein evolution.[13,52,53] While explicitly including epistatic interactions between all sites is computationally prohibitive, allowing for temporal variations in substitution processes (using a covarion-like process) and differences in preferences across sites (determined computationally or experimentally) are tractable ways of phenomenologically accounting for epistasis in inference models. Mutational effects, which appear inconsequential in experimental or computational settings may be exacerbated in nature. Further investigations into how nonadaptive processes alter evolutionary dynamics will be important, not only to better understand how proteins evolve, but also to better identify adaptive episodes when they occur in natural proteins.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Noor Youssef: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; software; validation; visualization; writing—original draft; writing-review & editing. **Edward Susko**: Conceptualization; methodology; supervision; writing-review & editing. **Andrew Roger**: Conceptualization; supervision; writing-review & editing. **Joseph Bielawski**: Conceptualization; resources; supervision; writing-review & editing.

GLOSSARY

Contingency	an increase in the fixation probability of a mutation due to preceding substitutions at other sites in the protein.
Convergence	the emergence of similar traits (phenotypic or genotypic) independently in multiple lineages.
Effective population size	a theoretical quantity conceptualized as the size of an idealized population (having an equal number of males and females, random mating, equal expectations of offspring for each individual, and a constant number of breeding individuals) exhibiting the same intensity of genetic drift as the natural population.
Entrenchment	a decrease in reversion rate due to subsequent substitutions at other sites in the protein.
Evolutionary anti-Stokes shift	a decrease in the propensity for a resident amino acid due to nonadaptive stability-mediated effects.
Evolutionary Stokes shift	an increase in the propensity for a resident amino acid due to nonadaptive stability-mediated effects.
Heterotachy	temporal variation in replacement rates through time.
Homoplasy	shared trait (phenotypic or genotypic) not due to presence in a common ancestor.
Homologous	shared trait (phenotypic or genotypic) due to presence in a common ancestor.
Random genetic drift	fluctuations in gene frequencies due to random sampling.
Reversions	a return to an ancestral state during evolution.
Selection coefficient	the selective advantage of a mutant measured as the difference in fitness effects of the mutant to wildtype.
ω rate ratio	inferred ratio of nonsynonymous to synonymous substitutions.

ORCID

Noor Youssef https://orcid.org/0000-0002-7708-982X

REFERENCES

1. Bazykin GA. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. Biol Lett. 2015;11:1–7.
2. Youssef N, Susko E, Roger A, Bielawski JP (2021). Trajectories of amino acid propensities under stability-mediated epistasis. Manuscript submitted for publication.
3. Pollock DD, Thiltgen G, Goldstein RA. Amino acid coevolution induces an evolutionary Stokes shift. Proc Natl Acad Sci. 2012; 109(21):E1352–E1359.
4. dos Reis M. How to calculate the non-synonymous to synonymous rate ratio of protein-coding genes under the fisher–wright mutation–selection framework. Biol Lett. 2015;11:20141031.
5. Bloom JD. Software for the analysis and visualization of deep mutational scanning data. BMC Bioinform. 2015;16:168.
6. Jones CT, Youssef N, Susko E, Bielawski JP. Shifting balance on a static mutation-selection landscape: a novel scenario of positive selection. Mol Biol Evol. 2017;34:391–407.
7. Echave J, Spielman S, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17: 109–121.
8. Youssef N, Susko E, Bielawski JP. Consequences of stability-induced epistasis for substitution rates. Mol Biol Evol. 2020; 37(11):3131–3148.
9. Wang HC, Li K, Susko E, Roger AJ. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC Evol Biol. 2008; 8:331.
10. Yang ZH, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 2002;19:908–917.

11. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005;22:2472–2479.
12. Goldstein RA, Pollock DD. Sequence entropy of folding and the absolute rate of amino acid substitutions. Nat Ecol Evol. 2017; 1:1923–1930.
13. Shah P, McCandlish DM, Plotkin JB. Contingency and entrenchment in protein evolution under purifying selection. Proc Natl Acad Sci. 2015;112(25):E3226–E3235.
14. Maynard Smith J. Natural selection and the concept of a protein space. Nature. 1970;225:563–564.
15. Goldstein RA. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. Genome Biol Evol. 2013;5:1584–1593.
16. Rodrigue N, Lartillot N. Detecting adaptation in protein-coding genes using a bayesian site-heterogeneous mutation-selection codon substitution model. Mol Biol Evol. 2017;34(1):204–214.
17. Bastolla U, Dehouck Y, Echave J. What evolution tells us about protein physics, and protein physics tells us about evolution. Curr Opin Struct Biol. 2017;42:59–66.
18. Teufel AI, Ritchie AM, Wilke CO, Liberles DA. Using the mutation-selection framework to characterize selection on protein sequences. Genes. 2018;9:409.
19. Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci. 2002;99:14878–14883.
20. Hochberg GKA, Liu Y, Marklund EG, Metzger BPH, Laganowsky A, Thornton JW. A hydrophobic ratchet entrenches molecular complexes. Nature. 2020;588:503–508.
21. Stern DL. The genetic causes of convergent evolution. Nat Rev Genet. 2013;14:751–764.
22. Parker J, Tsagkogeorga G, Cotton JA, et al. Genome-wide signatures of convergent evolution in echolocating mammals. Nature. 2013;502:228–231.
23. Mcgee MD, Wainwright PC. Convergent evolution as a generator of phenotypic diversity in threespine stickleback. Evolution. 2013;67:1204–1208.
24. Thomas GW, Hahn MW. Determining the null model for detecting adaptive convergence from genomic data: A case study using echolocating mammals. Mol Biol Evol. 2015;32:1232–1236.
25. Zou Z, Zhang J. No genome-wide protein sequence convergence for echolocation. Mol Biol Evol. 2015;32:1237–1241.
26. Goldstein RA, Pollard ST, Shah SD, Pollock DD. Nonadaptive amino acid convergence rates decrease over time. Mol Biol Evol. 2015;32(6):1373–1381.
27. Zou Z, Zhang J. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol Biol Evol. 2015;32:2085–2096.
28. Mendes FK, Hahn Y, Hahn MW. Gene tree discordance can generate patterns of diminishing convergence over time. Mol Biol Evol. 2016;33:3299–3307.
29. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian Protein Metabolism. New York: Academic Press, 1969; p. 21–132.
30. Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 2001;18:691–699.
31. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 1998;15:910–917.
32. Usmanova DR, Ferretti L, Povolotskaya IS, Vlasov PK, Kondrashov FA. A model of substitution trajectories in sequence space and long-term protein evolution. Mol Biol Evol. 2015;32:542–554.
33. Rokas A, Carroll S. Frequent and widespread parallel evolution of protein sequences. Mol Biol Evol. 2008;25:1943–1953.
34. Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. Epistasis as the primary factor in molecular evolution. Nature. 2012;490:535–538.
35. Naumenko SA, Kondrashov AS, Bazykin GA. Fitness conferred by replaced amino acids declines with time. Biol Lett. 2012;8: 825–828.
36. Muller HJ. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. Genetics. 1918;3:422–499.
37. Muller HJ. Reversibility in evolution considered from the standpoint of genetics. Biol Rev. 1939;14:261–280.
38. McCandlish DM, Shah P, Plotkin JB. Epistasis and the dynamics of reversion in molecular evolution. Genetics. 2016;203:1335–1351.
39. Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. Elife. 2013;2: e00631.
40. Popova AV, Safina KR, Ptushenko VV, et al. Allelespecific non-stationarity in evolution of influenza a virus surface proteins. Proc Natl Acad Sci. 2019;116:21104–21112.
41. Stolyarova A, Nabieva E, Ptushenko V, et al. Senescence and entrenchment in evolution of amino acid sites. Nat Commun. 2020;11:4603.
42. Gelbart M, Stern A. Site-specific evolutionary rate shifts in hiv-1 and siv. Viruses. 2020;12:1312.
43. Fowler DM, Fields S. Deep mutational scanning: A new style of protein science. Nat Methods. 2014;11:801–807.
44. Ashenberg O, Gong LI, Bloom JD. Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci. 2013;110(52):21071–21076.
45. Risso VA, Manssour-Triedo F, Delgado-Delgado A, et al. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. Mol Biol Evol. 2015;32:440–455.
46. Cherry JL. Should we expect substitution rate to depend on population size? Genetics. 1998;150:911–919.
47. Goldstein RA. The evolution and evolutionary consequences of marginal thermostability in proteins. Proteins. 2011;79:1396–1407.
48. Lunzer M, Golding B, Dean AM. Pervasive cryptic epistasis in molecular evolution. PLoS Genet. 2010;6:1–10.
49. Emlaw JR, Burkett KM, Dacosta CJB. Contingency between historical substitutions in the acetylcholine receptor pore. ACS Chem Nerosci. 2020;11:2861–2868.
50. Starr TN, Flynn JN, Mishra P, Bolon DNA, Thornton JW. Pervasive contingency and entrenchment in a billion years of hsp90 evolution. Proc Natl Acad Sci. 2018;115(17):4453–4458.
51. Fabien B, Roger AJ, Brown MW, Simpson AGB. The new tree of eukaryotes. Trends Ecol Evol. 2020;35(1):43–52.
52. Harms MJ, Thornton JW. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. Nature. 2014;512:203–207.
53. Sailer ZR, Thornton JW. High-order epistasis shapes evolutionary trajectories. Plos Comp Biol. 2017;13:e1005541.

54. Hietpas RT, Jensen JD, Bolon DNA. Experimental illumination of a fitness landscape. Proc Natl Acad Sci. 2011;108: 7896–7901.
55. Haddox HK, Dingens AS, Hilton SK, Overbaugh J, Bloom JD. Mapping mutational effects along the evolutionary landscape of hiv envelope. Elife. 2018;7:e34420.
56. Doud MB, Bloom JD. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. Viruses. 2016;8:155.
57. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 2020;16:e9380.
58. Doud MB, Ashenberg O, Bloom JD. Site-specific amino acid preferences are mostly conserved in two closely related protein homologs. Mol Biol Evol. 2015;32:2944–2960.
59. Lee JM, Huddleston J, Doud MB, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proc Natl Acad Sci. 2018;115: E8276–E8285.
60. Chan YH, Venev SV, Zeldovich KB, Matthews CR. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. Nat Comm. 2017;8:14614.
61. Bloom JD. An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol. 2014;31: 1956–1978.
62. Williams PD, Pollock DD, Blackburne BP, Goldstein RA. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comp Biol. 2006;2:e69.
63. Susko E, Roger AJ. Problems with estimation of ancestral frequencies under stationary models. Syst Biol. 2013;62(2): 330–338.
64. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How protein stability and new functions trade off. PLoS Comp Biol. 2008;4: 35–37.
65. Pollock DD, Goldstein RA. Strong evidence for protein epistasis, weak evidence against it. Proc Natl Acad Sci. 2014;111:2014.
66. Ferrada E. The site-specific amino acid preferences of homologous proteins depend on sequence divergence. Genome Biol Evol. 2019;11(1):121–135.
67. Yang Z, Wong WS, Nielsen R. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005;22:1107–1118.
68. Galtier N. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol. 2001;18:866–873.
69. Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP. Modeling the site-specific variation of selection patterns along lineages. Proc Natl Acad Sci. 2004;101:12957–12962.
70. Jones CT, Youssef N, Susko E, Bielawski JP. A phenotype-genotype codon model for detecting adaptive evolution. Syst Biol. 2020;69:722–738.
71. Bordner AJ, Mittelmann HD. A new formulation of protein evolutionary models that account for structural constraints. Mol Biol Evol. 2013;31(3):736–749.
72. Kleinman CL, Rodrigue N, Lartillot N, Philippe H. Statistical potentials for improved structurally constrained evolutionary models. Mol Biol Evol. 2010;27(7):1546–1560.
73. Rodrigue N, Kleinman CL, Philippe H, Lartillot N. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol Biol Evol. 2009;26(7):1663–1676.
74. Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL. Quantifying the impact of protein tertiary structure on molecular evolution. Mol Biol Evol. 2007;24:1769–1782.
75. Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 2003;20(10):1692–1704.
76. Bloom JD. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. Mol Biol Evol. 2014;31:2753–2769.
77. Hilton SK, Bloom JD. Modeling site-specific amino-acid preferences deepens phylogenetic estimates of viral sequence divergence. Virus Evol. 2018;4:vey033.
78. Spielman SJ, Wilke CO. The relationship between dn/ds and scaled selection coefficients. Mol Biol Evol. 2015;32:1097–1108.
79. Jones CT, Youssef N, Susko E, Bielawski JP. Phenomenological load on model parameters can lead to false biological conclusions. Mol Biol Evol. 2018;35:1473–1488.
80. Magee AF, Hilton SK, DeWitt WS (2020) Robustness of phylogenetic inference to model misspecification caused by pairwise epistasis. *bioRxiv*: 2020.11.17.387365.
81. Nasrallah CA, Mathews DH, Huelsenbeck JP. Quantifying the impact of dependent evolution among sites in phylogenetic inference. Syst Biol. 2011;60:60–73.
82. Kolaczkowski B, Thornton JW. A mixed branch length model of heterotachy improves phylogenetic accuracy. Mol Biol Evol. 2008;25:1054–1066.
83. Kolaczkowski B, Thornton JW. Long-branch attraction bias and inconsistency in bayesian phylogenetics. PLoS ONE. 2009; 4:e7891.
84. Whelan S. The genetic code can cause systematic bias in simple phylogenetic models. Philos Trans R Soc Lond B Biol Sci. 2008; 363:4003–4011.
85. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nat Methods. 2018;15:816–822.
86. Raimondi D, Tanyalcin I, FertCrossed JSD, et al. DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. Nucleic Acids Res. 2017;45:W201–W206.
87. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants From VarI-SIG 2014: Identification and annotation of genetic variants in the context of structure, function and disease. BMC Genomics. 2016;16:1–12.
88. Lu A, Guindon S. Performance of standard and stochastic branch-site models for detecting positive selection among coding sequences. Mol Biol Evol. 2014;31:484–495.

How to cite this article: Youssef N, Susko E, Roger AJ, Bielawski JP. Shifts in amino acid preferences as proteins evolve: A synthesis of experimental and theoretical work. Protein Science. 2021;30:2009–28. https://doi.org/10.1002/pro.4161