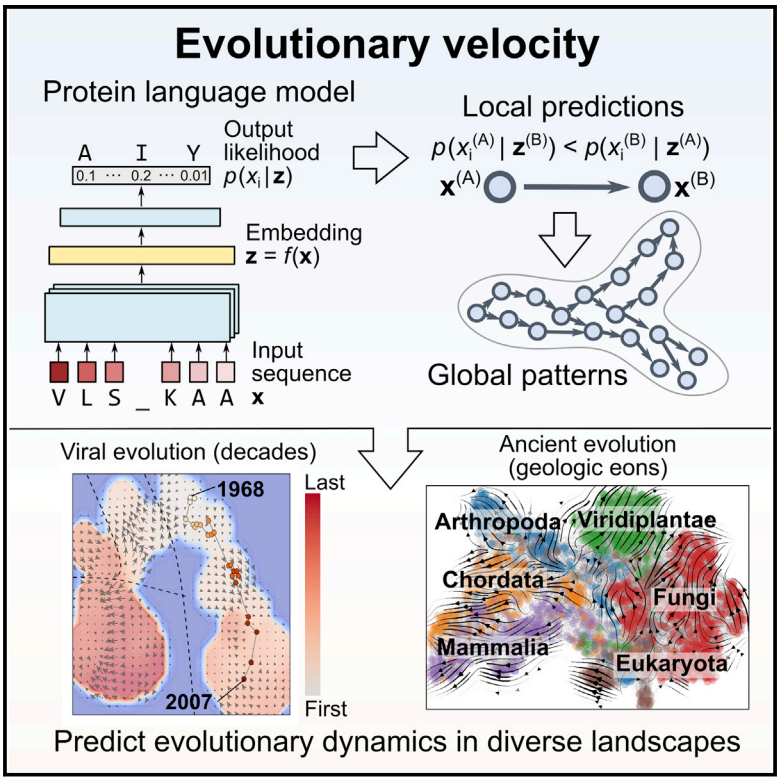
Cell Systems

Article

Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins

Graphical abstract



Highlights

- A single protein language model can predict mutational effects across diverse proteins
- Local predictions across a global sequence landscape can predict evolutionary order
- Model generalizes from evolutionary landscapes spanning years to geologic eons
- Protein language models learn general rules that aid evolutionary predictability

Authors

Brian L. Hie, Kevin K. Yang, Peter S. Kim

Correspondence

brianhie@stanford.edu (B.L.H.),
kimpeter@stanford.edu (P.S.K.)

In brief

Evolutionary patterns learned by a single protein language model can construct a "vector field" of protein evolution termed "evolutionary velocity" that predicts evolutionary dynamics across diverse proteins, from viral proteins evolving over years to highly conserved proteins evolving over geologic eons.

Hie et al., 2022, Cell Systems *13*, 274–285
April 20, 2022
https://doi.org/10.1016/j.cels.2022.01.003

CellPress

CellPress
OPEN ACCESS

Cell Systems

Article

Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins

Brian L. Hie,[1,2,5,*] Kevin K. Yang,[3] and Peter S. Kim[1,2,4,*]
[1]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305, USA
[2]Stanford ChEM-H, Stanford University, Stanford, CA 94305, USA
[3]Microsoft Research New England, Cambridge, MA 02142, USA
[4]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA
[5]Lead contact
*Correspondence: brianhie@stanford.edu (B.L.H.), kimpeter@stanford.edu (P.S.K.)
https://doi.org/10.1016/j.cels.2022.01.003

SUMMARY

The degree to which evolution is predictable is a fundamental question in biology. Previous attempts to predict the evolution of protein sequences have been limited to specific proteins and to small changes, such as single-residue mutations. Here, we demonstrate that by using a protein language model to predict the local evolution within protein families, we recover a dynamic "vector field" of protein evolution that we call evolutionary velocity (evo-velocity). Evo-velocity generalizes to evolution over vastly different timescales, from viral proteins evolving over years to eukaryotic proteins evolving over geologic eons, and can predict the evolutionary dynamics of proteins that were not used to develop the original model. Evo-velocity also yields new evolutionary insights by predicting strategies of viral-host immune escape, resolving conflicting theories on the evolution of serpins, and revealing a key role of horizontal gene transfer in the evolution of eukaryotic glycolysis.

INTRODUCTION

A longstanding open question in biology is whether evolution is predictable or fundamentally random (Gould, 1990; Lässig et al., 2017; Morris, 2003; de Visser and Krug, 2014). Theoretically, learning the rules that constrain evolution could enable some amount of evolutionary predictability. For example, mutating a protein site to a new residue that is biochemically incompatible with other residues could destabilize the protein, whereas a different new residue could instead improve stability and enable new types of mutations. However, biological complexity (for example, due to the combinatorial complexity of interactions among protein residues) makes learning these rules a considerable challenge (Bloom et al., 2006; Gong et al., 2013; Smith, 1970; Wright, 1932).

Promising advances in machine learning have improved the ability of a class of algorithms called language models to learn the rules that govern how amino acids can appear together to form a protein sequence (Alley et al., 2019; Bepler and Berger, 2019, 2021; Hie et al., 2021; Hsu et al., 2022; Rao et al., 2019; Rives et al., 2021; Madani et al., 2021). However, language models have been applied only to modeling local evolution, such as single-residue mutations, rather than more complex changes that occur over long evolutionary trajectories.

Here, we show how the evolutionary predictability enabled by a single, large language model (Box 1) provides a new method for recovering the dynamic trajectories of protein evolution that we refer to as "evolutionary velocity," or "evo-velocity." Evo-velocity is conceptually inspired by work in theoretical biology that understands evolution as a path that traverses a "fitness landscape" based on locally optimal decisions (Smith, 1970; Wright, 1932); although inspired by traditional fitness landscapes, evo-velocity is a distinct concept, as explained in the first paragraph of the discussion. Our key conceptual advance is that by learning the rules underlying *local* evolution, we can construct a *global* evolutionary "vector field" that we show can (1) predict the root (or potentially multiple roots) of observed evolutionary trajectories, (2) order protein sequences in evolutionary time, and (3) identify the mutational strategies that drive these trajectories. By predicting the order of biological sequences, evo-velocity has diverse applications that range from tracing the progression of viral outbreaks to understanding the history of life on earth.

In contrast to previous methods for predicting evolution, which assume protein-specific evolutionary models (for example, training a language model specific to a protein family) (Hie et al., 2021; Riesselman et al., 2018), we use a *single* language model to make all of our predictions across a diversity of proteins. We show that evo-velocity based on a single model generalizes to protein evolution across a breadth of organisms and evolutionary timescales—from the evolution of viral proteins over the course of years to the evolution of enzymes across all three domains of life—suggesting that algorithms can learn a common set of evolutionary rules, thereby expanding our ability to understand and predict protein evolution.

Box 1. Glossary

- Language model: a probability distribution over a sequence of tokens, for example, a sequence of English words or a sequence of amino-acid residues. Neural networks can implement a language model by outputting a probability value given an input sequence.
- Masked language model: a language model that predicts the identity of one or more masked tokens given all other tokens in the sequence. To train a neural network based on a masked language modeling objective, a subset of tokens in a sequence are masked and the language model is trained to predict the masked tokens based on the unmasked tokens.
- Language model pseudolikelihood: a score that approximates the likelihood of a full sequence but is computationally efficient to learn and compute. For example, the conditional likelihoods of single tokens learned by a masked language model can be used to efficiently compute a pseudolikelihood value for the full sequence; however, the joint likelihood of all tokens in the sequence is not directly learned by the masked language model.
- Sequence embedding: a computational representation of a sequence as a vector in a (typically binary or real-valued) vector space.
- Pseudotime: orders the elements of a set according to an inferred, one-dimensional value. When these elements correspond to the nodes in a graph, diffusion pseudotime is an algorithm that computes pseudotime based on the geodesic distance from a single, given "root" node.
- Fitness landscape: a comprehensive set of genotypes and their corresponding fitness values, where fitness is defined as a measure of the desirability of a given genotype. Fitness landscapes are often visualized by summarizing sequence variation in a one- or two-dimensional space and plotting fitness as the corresponding "height" over the space.
- Epistasis: a phenomenon where the effect of one mutation is dependent on the value of a different mutation. In statistical models of evolution where each genetic locus corresponds to a random variable, an epistatic model can account for statistical dependence among loci, whereas a non-epistatic model assumes independence among loci.
- Phylogenetic tree: a graph that describes evolutionary relationships in which nodes correspond to biological entities, edges connect evolutionarily related nodes, and the overall graph is a tree (there are no cycles). In a rooted phylogenetic tree, each edge is directed from an ancestral node to a descendant node, and each node is connected to a single ancestor expect for a single "root" node.
- Sequence similarity network: a generalization of a phylogenetic tree in which nodes correspond to biological sequences and edges connect evolutionarily related nodes but where cycles in the graph are allowed.

RESULTS

Overview of language models and evo-velocity

Our approach is conceptually inspired by the premise that evolution occurs through local changes that preserve or enhance evolutionary fitness (discussion) (Smith, 1970; Wright, 1932). In theory, predicting local evolution should, therefore, provide insight into global evolution as well (Figure 1A). To predict the local rules of evolution, we leverage protein language models, which learn the likelihood that a particular amino-acid residue appears within a given sequence context (Figure 1B). When trained on large corpuses of natural sequences, this language model likelihood is a strong correlate of the effects of mutations on laboratory measurements of protein fitness. For example, the ESM-1b (evolutionary scale modeling) language model (Rives et al., 2021), trained on ~27 million sequences in the UniRef50 database (Suzek et al., 2007; Table S1), can predict the effects of single-residue mutations as quantified by deep mutational scanning (DMS) of diverse proteins (Livesey and Marsh, 2020) (Figures 1C and S1; Data S1; STAR Methods). This correlation is comparable with that of a state-of-the-art mutational effect predictor (Riesselman et al., 2018) that was specially trained on sequence variation within individual protein families (Figure 1C); in contrast, ESM-1b is trained on a dataset that removes most intra-family sequence variation (Suzek et al., 2007). This broad predictive performance suggests that ESM-1b does not overfit to a single definition of fitness but learns general evolutionary patterns.

Our key hypothesis is that the likelihoods learned by these large-scale protein language models can be used to provide a notion of directionality within evolutionary trajectories. In our approach, which we call evo-velocity, we first model the "landscape" or the "manifold" (Yu et al., 2015) of sequence variation by constructing a sequence similarity network (Mccandlish, 2011) in which each node represents a protein sequence and edges connect similar sequences (Figure 1D). We quantify the sequence similarity as the Euclidean distance in language model-embedding space (Box 1), which can encode complex functional relationships (Bepler and Berger, 2019; Hie et al., 2021), and we construct the network by connecting a sequence to its *k*-nearest neighbors (KNN), which has been useful in modeling biological landscapes in many genomics applications (Becht et al., 2019; Hie et al., 2020; Wolf et al., 2018).

Then, language models assign a direction to each edge in the KNN network based on the change in the language model pseudolikelihood (Box 1) between the two sequences in that edge (Figure 1D). Intuitively, the local predictions of language models assign a "velocity" to pairs of sequences in the network that we assemble into an evolutionary "vector field" (La Manno et al., 2018). In this paper, we implement evo-velocity with a *single* masked language model, ESM-1b, but our framework can readily generalize to other implementations as well (discussion).

We can then determine whether there are consistent patterns in the "flow" of evolution across the global network (Figure 1D), which includes visualizing the trajectory and velocity in two dimensions (McInnes and Healy, 2018) and identifying the

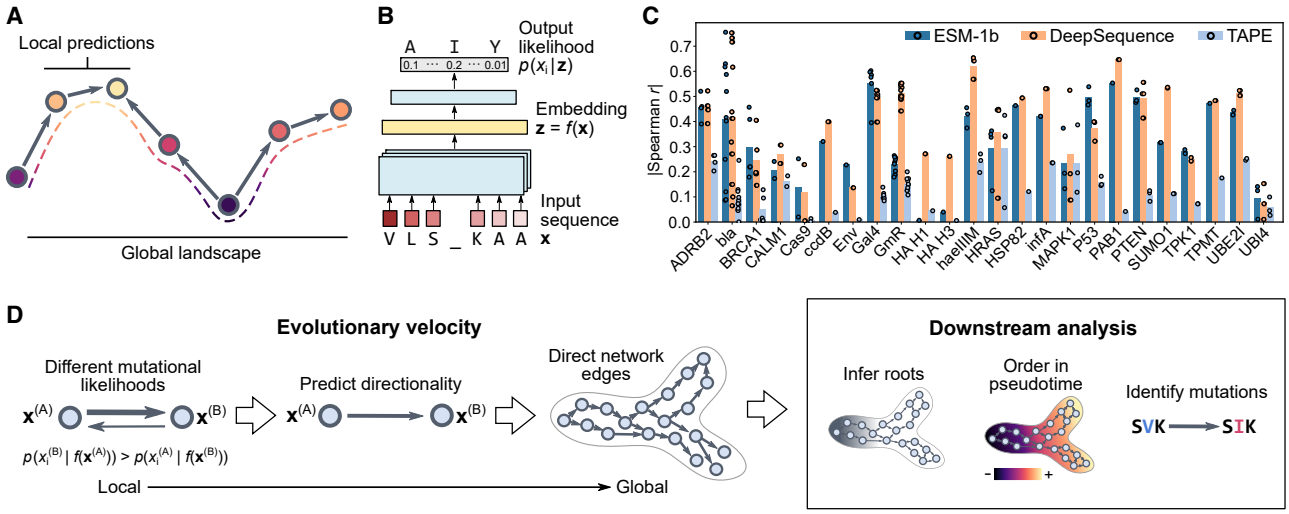


Figure 1. Constructing an evolutionary vector field by predicting local evolution

(A) A global evolutionary landscape can be approximated by a composition of local evolutionary predictions. In this cartoon, each circle corresponds to a sequence, where the proximity along the horizontal direction indicates sequence similarity and the height along the vertical axis represents "fitness."

(B) To make these predictions, we can leverage language models that learn the likelihood of an amino acid occurring within some sequence context.

(C) The pseudolikelihoods learned by language models correlate with the DMS-based measurements of various notions of protein fitness, without the language models being explicitly trained on these data (STAR Methods). Although DeepSequence trains a separate model for each protein family, ESM-1b and TAPE are general language models each trained on a single, non-redundant dataset. Circles indicate correlations of different DMS profiles within the same study (Data S1); bar height indicates the mean across these profiles.

(D) Evo-velocity uses language model likelihoods to assign a directionality to edges in a sequence similarity network, enabling downstream analysis, such as predicting root nodes, ordering nodes in pseudotime, and identifying the mutations associated with the largest changes in evo-velocity (STAR Methods).

mutations that correlate with the direction of evo-velocity. We can also use the language-model-inferred directionality to determine the roots of a diffusion process defined on the evo-velocity-weighted network. Once we have these roots, we can then place all sequences along a one-dimensional order, which we call diffusion pseudotime (Box 1), based on the distance along the trajectory from the estimated roots (Haghverdi et al., 2016). We hypothesize that pseudotemporal order reconstructs evolutionary order. We have provided a detailed methodology in STAR Methods.

Evo-velocity of influenza A nucleoprotein

As initial validation, we used evo-velocity to reconstruct the evolution of the nucleoprotein (NP) of influenza A virus. NP is an excellent evolutionary test case because its sequence evolution is densely sampled through influenza viral surveillance, and it undergoes natural selection in the form of host immune pressure but is less mutable than other viral proteins with a mutation rate of about one amino-acid residue per year (Gong et al., 2013). We obtained 3,304 complete NP sequences sampled from human hosts, constructed the sequence similarity network, and computed evo-velocity scores. When we visualized this network in two dimensions (McInnes and Healy, 2018), we observed phylogenetic structure corresponding to both the sampling year and influenza subtype (Figures 2A and S2A). The evo-velocity flow through the network (STAR Methods) corresponded to the known temporal evolution of NP (Figure 2A).

Because visualizing this flow in two dimensions can be prone to information loss or distortion through dimensionality reduction (La Manno et al., 2018), we sought to further quantify the relationship between evo-velocity and NP evolution. We first verified that, on average, the evo-velocity scores of the individual network edges increase along with greater differences in sampling time (Figure S2B). We then quantified global evo-velocity patterns using a diffusion analysis to estimate the network's roots (STAR Methods). We observed that the evo-velocity-inferred root sequences corresponded to the main species-crossover events in influenza history (Figure 2B), suggesting that our analysis accurately inferred the evolutionary origins of NP as observed in human hosts. We then used these roots to order sequences according to evo-velocity pseudotime (STAR Methods) and observed a significant correlation between the pseudotime and known sampling time (Spearman r = 0.49, two-sided *t*-distribution $p = 4 \times 10^{-197}$) (Figure 2C). We also observed that a well-characterized phylogenetic path of NP (Gong et al., 2013) progressed, on average, in the same direction as the evo-velocity gradient (Figures 2A and 2C) and agreed with simulated paths generated by random walks across our evo-velocity landscape (Figure 2D; STAR Methods). We emphasize that we did not provide our algorithm with any explicit knowledge of the known sampling time.

When comparing our evo-velocity landscape with a standard phylogenetic tree, we observed that evo-velocity can model more complex evolutionary relationships. For example, a midpoint-rooted phylogenetic tree of all NP sequences (STAR Methods) visually suggests that the H5N1- and H7N9-subtype sequences branch off from H1N1 (Figure 2E). Instead, evo-velocity predicts an independent origin of H5N1/H7N9 (Figures 2B and 2F), consistent with the epidemiological data indicating a recent zoonotic crossover of H5 and H7 avian influenza

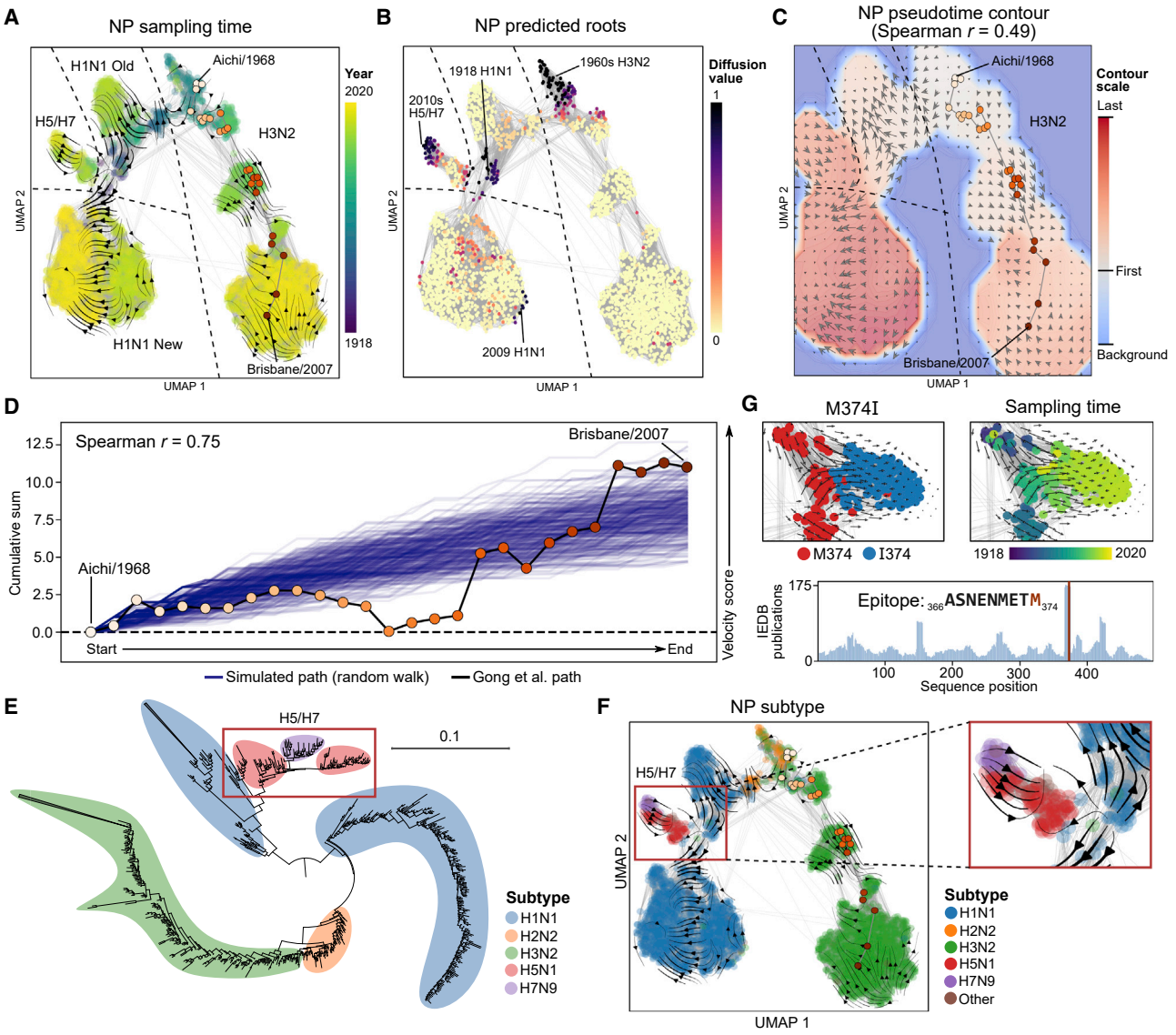


Figure 2. Evo-velocity of influenza A nucleoprotein

(A) The landscape of NP sequences (obtained from the Influenza Research Database, https://www.fludb.org/), represented as a KNN sequence similarity network, shows structure corresponding to the temporal evolution of various subtypes of influenza (Figure S2A); gray lines indicate network edges. Overlaying evo-velocity on the visualization as a streamplot shows a visual correlation between the flow of evo-velocity and known sampling time. A known phylogenetic path (orange circles) (Gong et al., 2013) starting with Aichi/1968 and ending with Brisbane/2007 moves in the direction of evo-velocity.

(B) Using the evo-velocity directionality to predict roots reveals four main root regions corresponding to the beginnings of different influenza pandemic events throughout history.

(C) Ordering sequences in pseudotime and visualizing pseudotime values in a two-dimensional contour plot shows pseudotime increasing in the direction of evo-velocity, which is visualized here as a two-dimensional field of evo-velocity vectors; contours correspond to pseudotime. Pseudotime and known sampling year have a Spearman correlation of 0.49 (two-sided *t*-distribution $p = 4 \times 10^{-197}$).

(D) On average, the Gong et al. path visualized in (A) and (C) has positive changes in evo-velocity scores over time and largely resembles simulated paths generated by performing random walks across our evo-velocity landscape (STAR Methods). A portion of the Gong et al. path with negative evo-velocity scores may be due to ordering ambiguities that are better resolved by considering evo-velocity.

(E) A maximum-likelihood, midpoint-rooted phylogenetic tree of all NP sequences conveys that the H5N1 and H7N9 subtype sequences branch off from H1N1 sequences.

(F) In contrast, evo-velocity predicts an independent origin of H5N1/H7N9 influenza (Sutton, 2018) (see B) and sequence similarity with H1N1 due to convergent evolution.

(G) The M374I mutation to NP has the second strongest magnitude change in evo-velocity (STAR Methods) and is located in the most well-studied human T cell epitope on NP (Table S2).

(Sutton, 2018). Evo-velocity also predicts that the observed similarity of H5N1/H7N9 and H1N1 NP sequences sampled in human hosts is due to convergent evolution (Figure 2F), which is challenging to explicitly represent using a phylogenetic tree.

We next sought to use our evo-velocity landscape to provide new insight into NP evolution. We, therefore, identified the mutations that corresponded to the strongest changes in the evo-velocity scores (STAR Methods). Of the top five such mutations in NP, all are present in experimentally validated human T cell epitopes, and one of these mutations, M374I, is located in the most well-characterized linear NP epitope in the Immune Epitope Research Database (IEDB) (Vita et al., 2015; Figures 2G and S2C; Table S2). Moreover, all five mutations involve a single-nucleotide substitution resulting in a methionine that changed to a hydrophobic or polar-uncharged amino-acid residue, suggesting a possible T cell escape strategy that has recurred in multiple NP epitopes throughout history (Figures 2G and S2C; Table S2).

All NP sequences in our analysis belong to a single UniRef50 sequence cluster (Suzek et al., 2007) for which the representative sequence is from a 1934 H1N1 virus (Figure S2D). We found that a similarity to sequences present in UniRef50, the ESM-1b training dataset, does not explain evo-velocity pseudotime (Table S3; STAR Methods) and that evo-velocity pseudotime was not explained by variation in sequence length (Table S4). We also found that computing evo-velocity scores with a smaller language model, the TAPE (tasks assessing protein embeddings) transformer model (Rao et al., 2019), trained with a different model architecture on the Pfam database of protein families (El-Gebali et al., 2019), closely reproduced the ESM-1b evo-velocity results (Spearman $r = 0.93$, two-sided t-distribution $p < 1 \times 10^{-308}$) (Tables S5 and S6; Figures S2E and S2F). Together, these results suggest that our evo-velocity results are not explained by a trivial language model preference to UniRef50.

Evo-velocity was, therefore, able to reconstruct the direction of NP evolution without any explicit knowledge of influenza subtype or sampling time. Moreover, we found that the generic rules learned by large language models were sufficient to predict the evolution of a specific protein.

Evo-velocity of viral proteins

Given the promising results for NP, we were interested in seeing if evo-velocity could generalize to other viral proteins as well. We next analyzed the evolution of influenza A hemagglutinin (HA), a more variable protein on the viral surface responsible for viral-host membrane fusion (Eckert and Kim, 2001; Harrison, 2008). The HA sequence landscape contains two main trajectories, one beginning in 1918 and the other beginning in 2009 (Figure 3A). We observed that the 2009-rooted trajectory first becomes more similar (i.e., more proximal in sequence-embedding space) to the 1918-pandemic HA before subsequent divergence (Figure 3A), consistent with the convergent antigenic similarity between 1918 and 2009 pandemic influenza (Wei et al., 2010; Xu et al., 2010). As with NP, an evo-velocity analysis of 8,115 HA sequences recovered roots corresponding to the known origins of HA H1 in humans from the 1918 and 2009 H1N1 pandemics, and evo-velocity pseudotime was strongly correlated with sampling date (Spearman $r = 0.51$, two-sided t-distribution $p < 1 \times 10^{-308}$) (Figures 3A and 3B). Despite the sequence variability of HA being higher than that of NP, evo-velocity was still able to reconstruct the trajectory and directionality of HA evolution.

As with NP, our HA pseudotime results were not explained by sequence similarity to the training dataset (Figure S3A; Table S3). We were also able to use TAPE-based velocities to identify similar root regions in the post-2009 pandemic trajectory, but TAPE had a more difficult time identifying the 1918 sequences as oldest, most likely due to TAPE's smaller model size and less capable mutational effect predictions (Figures 1C and S3B–S3D; Table S5).

We next analyzed the evolution of the group-specific antigen (Gag) polyprotein of human immunodeficiency virus type 1 (HIV-1) using 18,018 sequences. Visualizing the sequence similarity network overlaid with evo-velocity reveals a flow corresponding to the known subtype branching history of HIV-1, with circulating recombinant forms (for example, subtypes AE and BC) branching off of the main subtypes and occurring later in pseudotime (Figures 3C and 3D). HIV-1 Gag sequences also had strong positive velocities compared with phylogenetically similar Gag sequences from chimpanzee simian immunodeficiency virus (SIVcpz) (Figure S3E), consistent with a SIVcpz origin preceding the evolution of pandemic HIV-1 (Sharp and Hahn, 2011). We observed weaker correlation between pseudotime and sampling date (Spearman $r = 0.12$, two-sided t-distribution $p = 6 \times 10^{-49}$) (Figure S3F) compared with influenza proteins, consistent with weak population-level immune pressure on Gag evolution. Gag pseudotime was not explained by sequence similarity to UniRef50 (Table S3) and was also reproducible using TAPE-based velocities (Figure S3G; Table S5).

We next applied our algorithm to analyze 75,584 sequences of the spike glycoprotein of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) across a much shorter historical timescale of around 21 months. The sequence similarity network reconstructs the overall trajectory of spike evolution, and evo-velocity analysis identifies the sequence clusters associated with later sequences, including the B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.1 (Kappa), B.1.617.2 (Delta), and P.1 (Gamma) variants-of-concern (Maher et al., 2022; Walensky et al., 2021), as later in pseudotime (Figures 3E–3G). Despite a shorter evolutionary timescale, evo-velocity pseudotime and sampling date still had a Spearman correlation of 0.55 (two-sided t-distribution $p < 1 \times 10^{-308}$). We also note that SARS-CoV-2 spike evolution occurred outside of the temporal range associated with both language model training datasets, and we were also able to reproduce the results with TAPE-based evo-velocity (Figure S3H; Table S5).

Across these four viral proteins, therefore, evo-velocity was able to reconstruct the directionality of evolution, consistent with the known trajectories. All of our analysis was based on a single model that was trained without any explicit knowledge of viral sampling date, subtype, or protein-specific sequence variation.

Evo-velocity of eukaryotic proteins

After validating our approach with known viral trajectories, we wanted to see if evo-velocity could generalize to longer

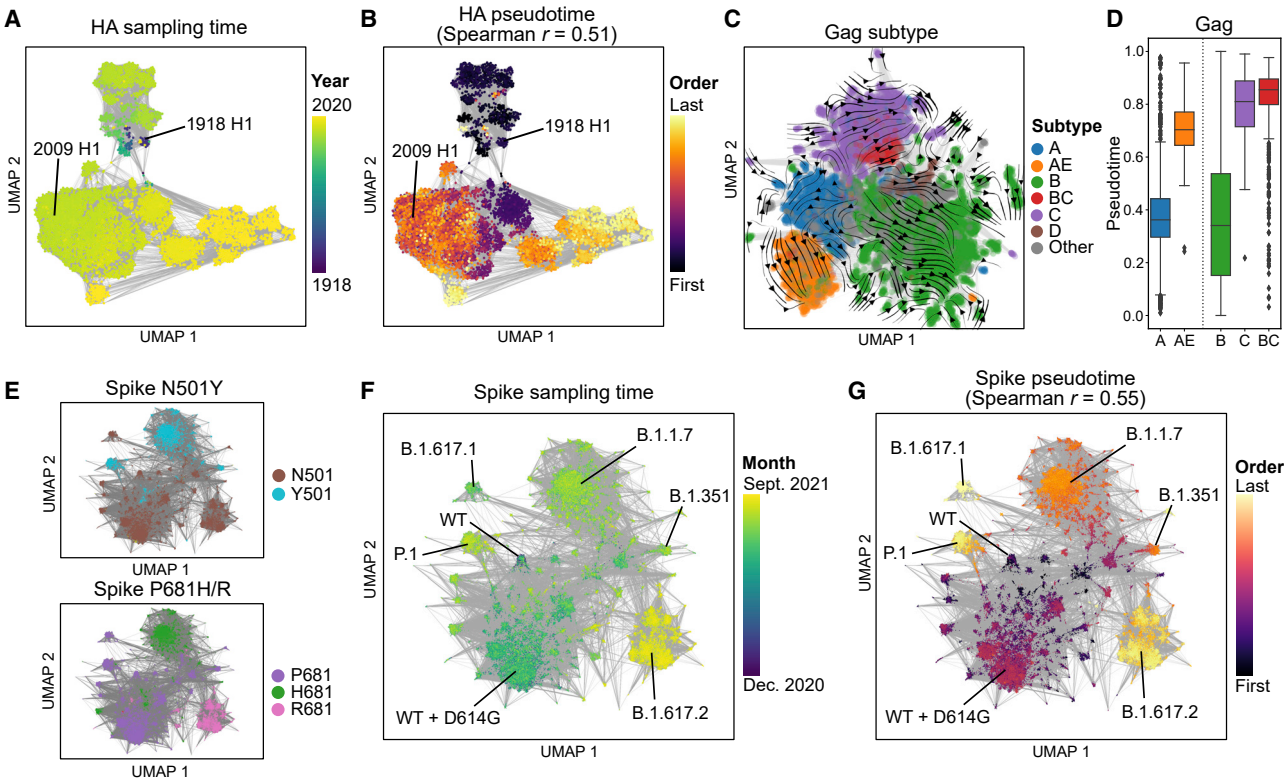


Figure 3. Evo-velocity of viral proteins

(A and B) Temporal evolution of HA H1 sequences (obtained from the Influenza Research Database, https://www.fludb.org/) is captured in the UMAP landscape and is also predicted by evo-velocity pseudotime. Two main clusters correspond to the two main pandemic trajectories of H1N1, the first beginning in the early twentieth century and the second beginning in the early twenty-first century, with some convergence between the two trajectories consistent with the known antigenic convergence (Wei et al., 2010; Xu et al., 2010). Pseudotime and known sampling time have a Spearman correlation of 0.51 (two-sided *t*-distribution $p < 1 \times 10^{-308}$).

(C and D) An evo-velocity streamplot of Gag evolution illustrates the branching trajectories of HIV-1 subtypes, including major subtypes, such as A, B, and C, preceding circulating recombinant forms, such as AE and BC. Box extends from the first to third quartile with a line at the median; whiskers extend to 1.5 times the interquartile range, and diamonds indicate outlier points. Gag sequences were obtained from the Los Alamos National Laboratory (LANL) HIV database (https://www.hiv.lanl.gov/).

(E–G) Variants of spike (identified using characteristic mutations, such as D614G and N501Y) that emerge in later portions of the COVID-19 pandemic are also predicted to be later in evo-velocity pseudotime. Pseudotime and known sampling time have a Spearman correlation of 0.55 (two-sided *t*-distribution $p < 1 \times 10^{-308}$).

trajectories, such as protein evolution that spans multiple species. Although we have access only to extant sequences, we hypothesized that evo-velocity might still provide useful orderings if some extant sequences are closer to the ancestral sequence than others. As an initial test case, we analyzed the globin protein family because of its extensive phylogenetic characterization (Pillai et al., 2020), including laboratory reconstruction of ancestral intermediates, that we can use to validate our model (Figure 4A).

The landscape of 6,097 eukaryotic globin sequences forms a branching trajectory with three major divisions corresponding to myoglobin, alpha hemoglobin, and beta hemoglobin (Figure 4B). The predicted root region lies in the part of the landscape closest to neuroglobin and cytoglobin (Figures 4B, S4A, and S4B). Of the major classes of globins, neuroglobin is estimated to be occurring earliest in pseudotime, whereas the alpha (Hbα) and beta (Hbβ) subunits of hemoglobin occur last in pseudotime (Figure 4C), consistent with a previous analysis of globin phylogeny by Pillai et al. (Figure 4A). These results are also reproducible when using TAPE to compute the evo-velocity scores (Figures S4C and S4D; Table S5) and when controlling for sequence similarity to the training dataset (Figure S4D; Table S3; STAR Methods).

Previous work (Pillai et al., 2020) also reconstructs ancestral globins that are confirmed to be viable oxygen binders and progress from a monomeric myoglobin/hemoglobin ancestor (AncMH) to a dimeric alpha/beta hemoglobin ancestor (Ancα/β) to a tetramer formed by separate alpha and beta hemoglobin ancestors (Ancα and Ancβ, respectively) (Figure 4A). Consistent with evo-velocity increasing over evolutionary time, the ESM-1b language model likelihood, on average, increases from AncMH to extant myoglobin and hemoglobin sequences, but this improvement diminishes for more proximal ancestors (Figure S4E). Together, our globin results suggest that the evo-velocity pseudotime within a protein family can recover ordering relationships over longer evolutionary timescales.

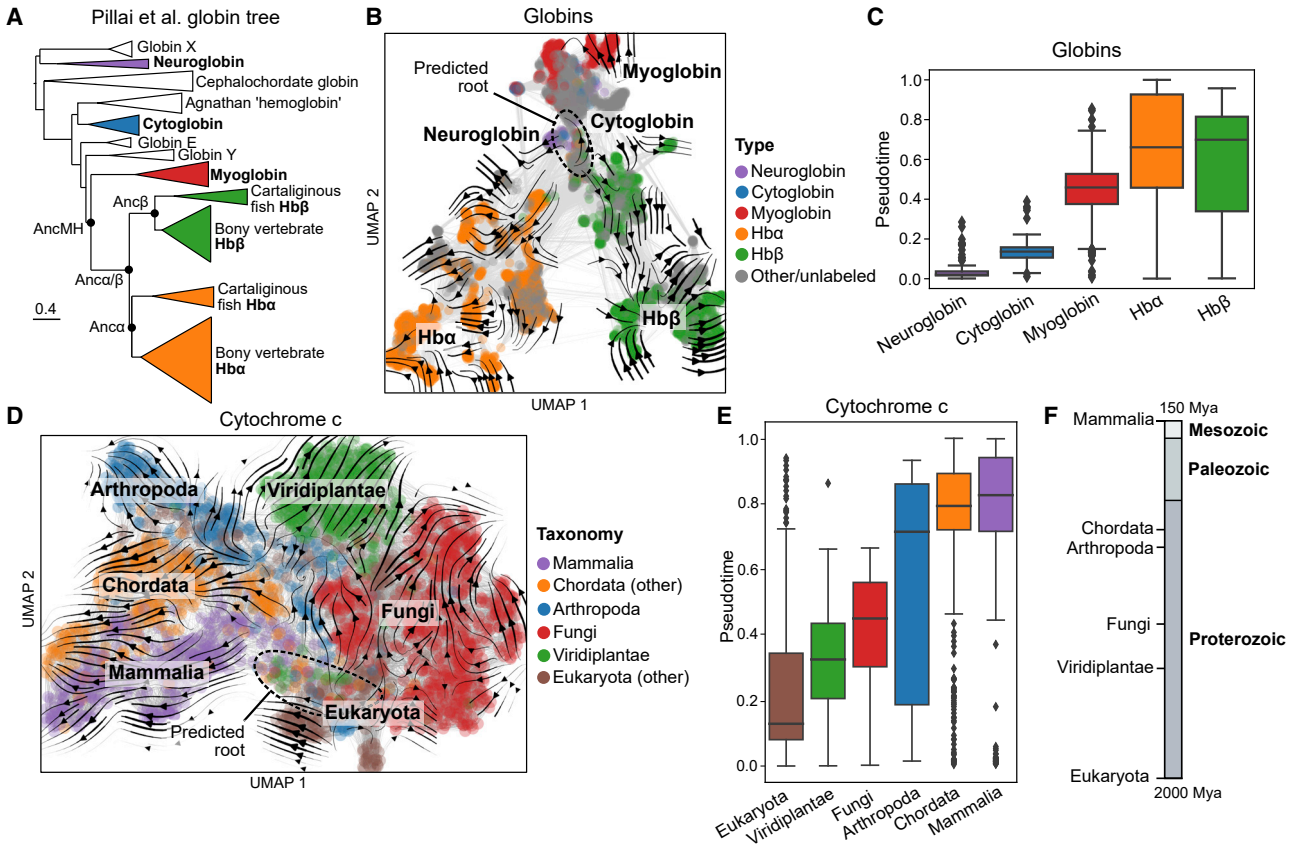


Figure 4. Evo-velocity of eukaryotic proteins

(A) The maximum likelihood phylogenetic tree determined by Pillai et al. (2020) is rooted in globin X and neuroglobin with the longest branches extending to Hbα and Hbβ.

(B) The landscape of globin sequences (obtained from UniProt, https://www.uniprot.org/) shows a branching trajectory with the predicted root also closest to neuroglobin (Figure S4A).

(C) Computing pseudotime from this predicted root places Hbα and Hbβ as most recent in evolution, consistent with the tree of Pillai et al.

(D) The landscape of cytochrome c sequences (obtained from UniProt, https://www.uniprot.org/) shows clustering structure corresponding to the known taxonomic labels, with the evo-velocity gradient beginning among single-celled eukaryotes and plants (Figure S5A).

(E and F) The ordering of the median evo-velocity pseudotimes of various taxonomic labels corresponds to the evolutionary orderings in geologic time determined by molecular clocks and the fossil record (Hedges et al., 2015). For all boxplots: box extends the from first to third quartile with a line at the median; whiskers extend to 1.5 times the interquartile range, and diamonds indicate outlier points.

To further test this hypothesis, we analyzed 2,128 sequences of cytochrome c, a well-studied protein in evolutionary biology because of its high sequence conservation among most eukaryotes (McLaughlin and Dayhoff, 1973). When visualized, the sequence similarity network combined with evo-velocity reflects the taxonomic diversification of the eukaryota (Figure 4D). The ordering of the median pseudotimes of different taxonomic classes also recapitulates their known ordering in geologic time based on estimates from the fossil record and molecular clocks (Hedges et al., 2015) (Figures 4E, 4F, S5A, and S5B; Table S6). We were also able to reproduce pseudotemporal orderings when using TAPE to compute the evo-velocity scores (Figures S5C and S5D; Tables S5 and S6) and when controlling for sequence similarity to the training dataset (Figure S5D; Tables S3 and S6). In total, therefore, our analysis of well-studied eukaryotic protein families demonstrates that evo-velocity can generalize to protein evolution at much longer timescales.

Evo-velocity of ancient evolution

After validating that evo-velocity could reconstruct longer trajectories of protein evolution, we applied evo-velocity to highly conserved proteins, which often have substantial evolutionary uncertainty (Weiss et al., 2016), to yield new insight into ancient evolution. A protein family with considerable evolutionary uncertainty is that of the serine protease inhibitors, or serpins (Irving et al., 2002; Roberts et al., 2004). Unlike most highly conserved families in which most of the diversity is bacterial, most of the diversity among serpins is eukaryotic, which we likewise observe in our landscape of 22,737 serpin sequences (Figures 5A and 5B). This has led to conflicting theories as to whether serpins indeed have a phylogenetic root in eukaryotes, with prokaryotes acquiring serpins via horizontal gene transfer (HGT), or if this root is an artifact of a greater eukaryotic diversity biasing phylogenetic root estimation (Irving et al., 2002; Roberts et al., 2004; Spence et al., 2021). Since evo-velocity is not prone to the same bias when estimating roots, we used evo-velocity to order

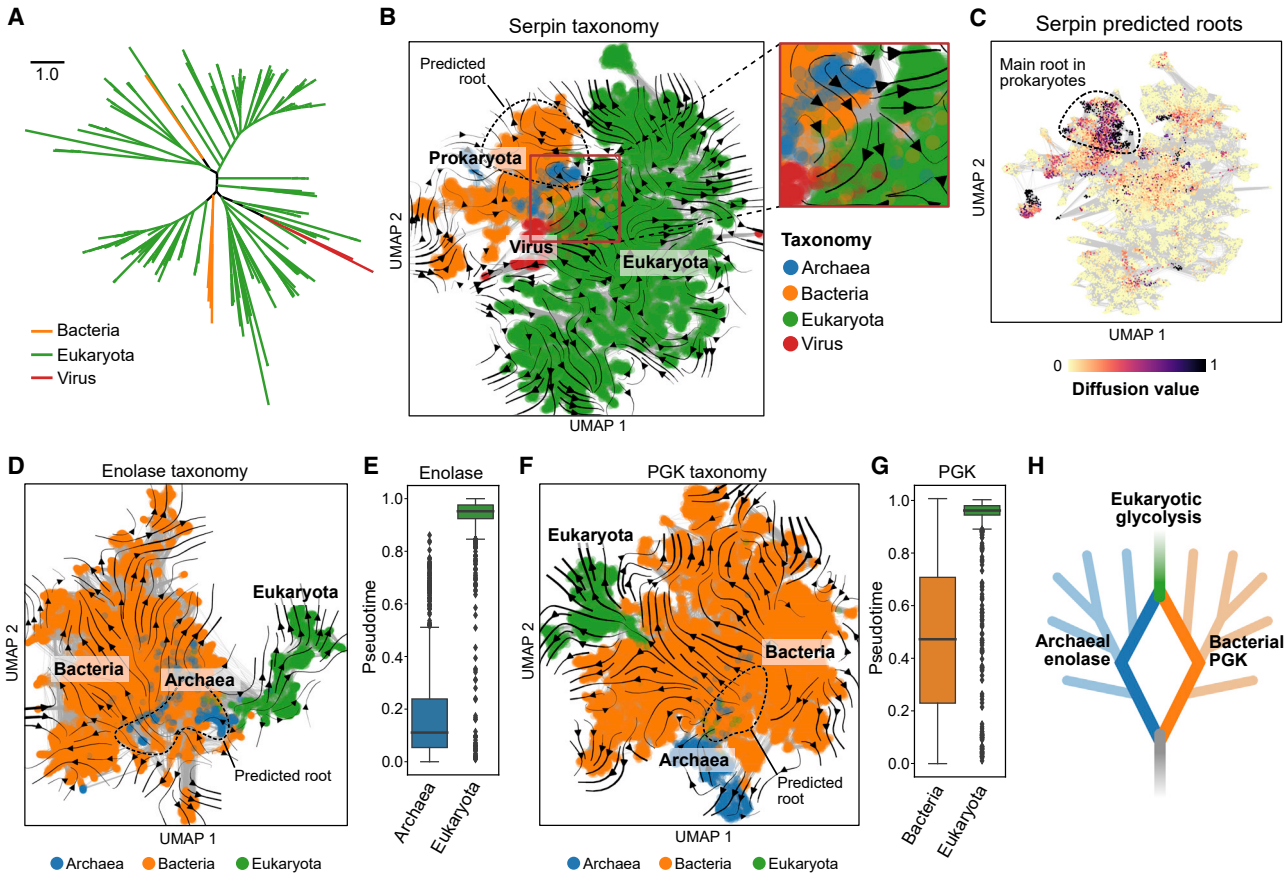


Figure 5. Evo-velocity of ancient evolution

(A) The unrooted maximum likelihood phylogenetic tree of serpins shows substantially more eukaryotic than prokaryotic diversity, leading some to hypothesize a eukaryotic root (Irving et al., 2002; Roberts et al., 2004).

(B and C) Despite lower prokaryotic diversity, evo-velocity still identifies the root of serpins within the prokaryotes, and eukaryotes are the last domain in evo-velocity pseudotime (Figure S6A), suggesting that prokaryotic serpins were not acquired from eukaryotes via HGT (Irving et al., 2002; Roberts et al., 2004). Serpin sequences were obtained from UniProt (https://www.uniprot.org/).

(D and E) The evo-velocity-predicted root of the enolase landscape begins in a region of archaea and some bacteria, with eukaryotic enolase having the highest pseudotime and being directly proximal to archaeal enolase on the sequence landscape (Figures S6B, S6C, and S6F). Enolase sequences were obtained from UniProt (https://www.uniprot.org/).

(F and G) The evo-velocity-predicted root of the PGK landscape begins in a mostly bacterial region with some archaea, with eukaryotic PGK also being the highest in pseudotime and directly proximal to bacterial PGK (Figures S6D, S6E, and S6G). PGK sequences were obtained from UniProt (https://www.uniprot.org/).

(H) The sequence landscapes and evo-velocity-predicted roots suggest that the component enzymes of eukaryotic glycolysis were acquired through different evolutionary paths via HGT; figure adapted from Figure 1 of Weiss et al. (2016). For all boxplots, box extends from the first to third quartile with a line at the median, whiskers extend to 1.5 times the interquartile range, and diamonds indicate outlier points.

serpin sequences in pseudotime and found that the main predicted root region was located among the prokaryotes (Figures 5B, 5C, and S6A). These results, along with the uncertain mechanism of eukaryotic-to-prokaryotic HGT (Spence et al., 2021), provide strong evidence that serpin evolution follows a more canonical trajectory. The evo-velocity landscape of serpin evolution also contains a class of viral serpins that are predicted to have evolved from eukaryotic serpins (Figure 5B), which is consistent with the viral repurposing of mammalian host serpins as previously hypothesized (Chen et al., 2011).

We next analyzed two of the most conserved glycolytic enzymes, enolase and phosphoglycerate kinase (PGK) (Piast et al., 2005; Potter and Fothergill-Gilmore, 1993; Rojas-Pirela et al., 2020). The landscape of 31,901 sequences from the enolase family shows a clear evo-velocity-predicted root region located in bacterial and archaeal sequences (Figures 5D, S6B, and S6C). Archaea are also oldest in pseudotime and eukaryota are newest, with bacteria showing considerable pseudotemporal variation (Figures 5E and S6C). The landscape of 30,455 PGK sequences has a similar origin in a region with bacterial and archaeal sequences (Figures 5F, S6D, and S6E), although with more pseudotemporal variation among archaeal PGK (Figures 5G and S6E).

The largest difference between the enolase and PGK landscapes lies in the location of eukaryota: although both landscapes place eukaryota as higher in pseudotime, eukaryotic enolase sequences branch off of archaeal enolase but eukaryotic PGK sequences branch off of bacterial PGK (Figures 5D

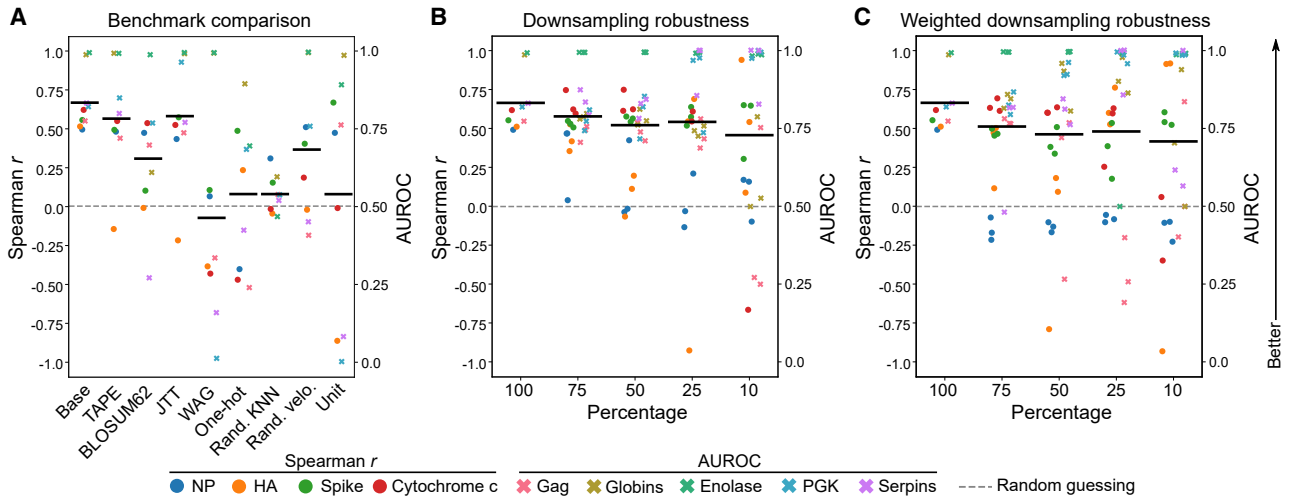


Figure 6. Evo-velocity benchmarks

(A) Metrics quantifying predicted evolutionary order were computed for the model based on ESM-1b embeddings and velocities (Base); velocities instead computed by TAPE, blocks of amino acid substitution matrix-62 (BLOSUM62), Jones-Taylor-Thornton (JTT), or Whelan and Goldman (WAG) evolutionary models; embeddings based on a one-hot binary encoding (one-hot); a sequence landscape with randomly assigned edges (Rand. KNN); velocities computed by randomly initialized ESM-1b (Rand. velo.); or unit velocities (STAR Methods).

(B and C) We also computed the same metrics after applying the base model to sequence landscapes that were downsampled to the percentage of the original dataset indicated on the horizontal axis, repeating across three random seeds. Downsampling was either done uniformly (B) or weighted to preferentially remove older sequences (C). In all plots, metrics based on correlation (Spearman *r*) or separation (area under the receiver-operating characteristic curve [AUROC]) were scaled to occupy the same vertical axis. Black horizontal lines indicate the mean within each category.

and 5F); similar patterns were also observed when visualizing the unrooted phylogenetic trees of both proteins (Figures S6F and S6G). These results suggest an archaeal origin of eukaryotic enolase and a bacterial origin of eukaryotic PGK (Figure 5H) and are consistent with HGT contributing to a mixture of archaeal and bacterial genes in the last eukaryotic common ancestor (Weiss et al., 2016). These results are also consistent with a component-wise evolution of glycolysis (Potter and Fothergill-Gilmore, 1993), rather than the pathway being inherited in totality from a single organism.

In all the three highly conserved proteins that we tested, we were able to reproduce evo-velocity pseudotime even when explicitly controlling for sequence similarity to the training dataset (Figures S6A and S6H; Tables S3 and S6) and when using TAPE to compute the evo-velocity scores (Figures S6A and S6H; Tables S5 and S6). The variability in sequence length did not explain evo-velocity pseudotime (Table S4). Moreover, the direction of the evo-velocity gradient is not explained by a trivial training set bias toward eukaryotes, as most of the sequences in UniRef50 are bacterial (Table S1), and we emphasize that no explicit taxonomic information was provided to our algorithm. Rather, our results suggest that evo-velocity can provide insight into evolution at the longest evolutionary timescales.

Evo-velocity benchmarks

Given the ability of evo-velocity to predict the directionality of evolution, we wanted to assess how the individual components of our algorithm contribute to overall performance. We performed control experiments in which we implemented sequence representation, KNN network construction, or velocity score computation with a simpler alternative (STAR Methods). We found that our base model, using ESM-1b embeddings and velocity scores, consistently outperformed all benchmark methods (Figure 6; Table S6; STAR Methods).

A notable control experiment demonstrated that velocities based on a JTT amino-acid substitution matrix (Jones et al., 1992), a non-epistatic evolutionary model, still had reasonable overall performance (Figure 6A; Table S6). This experiment also helped us reason about the contributions of modeling drift (via substitution matrices that consider the similarity of two residues) and, additionally, epistasis (via masked language models that use the entire sequence context to predict mutational likelihood). For proteins under weak selection, such as HIV-1 Gag, both JTT and the epistatic models (i.e., ESM-1b and TAPE) have comparable predictive ability (Table S6); however, for proteins under strong selection, such as influenza A HA, only the largest epistatic model, ESM-1b, can predict the direction of evolution (Table S6), consistent with previous observations that stronger selection promotes epistatic interactions (Gupta and Adami, 2016; Hayden and Wagner, 2012). Moreover, JTT has less consistent predictions of evolutionary direction than the epistatic models when visualizing evo-velocity (Figure S7), indicating that epistatic models enable stronger, more consistent evolutionary predictions.

The other control experiments revealed that using evolutionary information to calculate both embeddings and velocities is crucial to performance, especially when generalizing across diverse proteins; for example, diffusion analysis based on unit velocity, similar to midpoint-rooted phylogenies, places the root of serpins among the more diverse eukaryotic sequences (Table S6). Our control experiments also revealed robustness to missing sequences, even when removing as much as 75% of the initial landscape (Figure 6; Data S2; STAR Methods). In

total, these results support our overall design choices, including landscapes based on neural embeddings and likelihoods based on a large, epistatic model.

DISCUSSION

Relationship to protein fitness landscapes

Although conceptually inspired by landscape-based evolutionary theory, evo-velocity departs from traditional notions of a fitness landscape in a few ways. First, although we show that language model likelihood correlates with the laboratory measurements of mutational effects (Figure 1C), protein "fitness" in nature is much more complex and dynamic (although we note that ESM-1b is trained only on diverse, natural sequence variation, making it more difficult for the model to be biased toward a single notion of fitness). Second, because we define local transitions in the evo-velocity landscape as probabilities, the most probable outgoing edge might still correspond to a negative velocity score (due to lower language model likelihood) (Figures S8A and S8B). Third, network diffusion is predominately influenced by average-case patterns and can tolerate local inconsistencies in the direction of velocity (for example, due to noise or occasional decreases in fitness) (Figure S8C). For these reasons, a sequence with high pseudotime may not correspond to one with high likelihood or high "fitness" (Figure S8C). Rather, our study finds that the diffusion pseudotime computed on the landscapes of natural protein families is most predictive of evolutionary order.

Limitations of the study

Our study is limited to observed sequences, especially those deposited in public databases. Because the actual evolutionary order of these observed sequences is fundamentally unknown, we instead must rely on known sampling times or taxonomic metadata for orthogonal validation. Our benchmarking experiments also indicate that robust estimates of evolutionary order require at least a few hundred evolutionarily related sequences. Using a masked language model to predict mutational effects is not well defined in the case of insertions or deletions, which we currently do not consider when computing velocity scores (STAR Methods). Although this study closely analyzes a diverse set of proteins, we did not perform a more exhaustive analysis (for example, across thousands of protein families), which could reveal outlier examples with potentially interesting evolutionary histories.

Additional discussion and future directions

Here, we show that large-scale protein language models can learn evolutionary rules well enough to predict the directionality of evolution. The sufficiency of a single, large protein language model to predict the evolutionary dynamics of diverse proteins implies that there are common rules constraining natural evolution and that language models can extract these rules from data. Because ESM-1b is trained on sequence alone, these rules most likely correspond to intrinsic, fundamental properties of proteins, such as stability or evolvability. A language model could also indirectly learn some environmental information that is encoded within sequence variation (Hie et al., 2021).

Evo-velocity has a number of distinctives with respect to phylogenetic tree reconstruction. Evo-velocity is especially suitable for analyzing large (~1,000 or more) collections of sequences. We currently limit our analysis to extant sequences, rather than artificially reconstructing ancestral sequences, although these could be incorporated into the analysis as well. Evo-velocity also admits multiple roots that are better mathematically determined than phylogenetic roots (Kim et al., 2020; Masuda et al., 2017) (although users can manually specify root sequences as well). Evo-velocity landscapes can also model convergent evolution (Figure 2F), in contrast to the divergence assumption that underlies reconstructing phylogenetic trees.

We also find that evo-velocity provides a helpful notion of uncertainty in its predictions that is less natural to obtain from standard phylogenetic methods. For example, evo-velocity reports multiple roots, indicating evolutionary ambiguity regarding the oldest sequences or reflecting discontinuous trajectories due to missing evolutionary ancestors. Similarly, the most robust ordering relationships are at the level of groups of sequences, whereas variation in the diffusion pseudotime within a taxonomic group can reflect the uncertainty of a given ordering prediction.

Our experiments also reveal that more complex epistatic models improve both predictive performance and generality (Figure 6), especially for proteins (such as influenza A HA) undergoing strong selection. These findings raise a number of interesting questions, including the degree to which the rules learned by language models are biologically interpretable, for example, in terms of thermostability or evolvability (Bloom et al., 2006; Gong et al., 2013). Another question is whether larger protein language models, fine-tuning (Luo et al., 2021) or directly using *in-vitro* assays to compute velocity scores could improve the performance, resolution, and interpretability of evo-velocity.

Promisingly, evo-velocity offers a new approach through which to reevaluate current evolutionary hypotheses. For example, when evaluating a potential hypothesis of eukaryote-to-prokaryote HGT among serpins (Irving et al., 2002; Roberts et al., 2004), evo-velocity instead predicted a more canonical evolutionary trajectory (Figure 5; Spence et al., 2021). Although we mostly take a gene-centric approach to evolution (Dawkins, 1976), trajectories could also be integrated across multiple genes to provide insight into evolution at the level of pathways (as done for our analysis of glycolytic enzymes), gene modules, or even whole genomes. This might enable the calibration of evo-velocity pseudotime to historical or geologic time (which may have a non-linear relationship), providing an additional method for dating evolutionary events. Evo-velocity also suggests a way to predict future evolution and to design novel protein sequences.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2022.01.003.

ACKNOWLEDGMENTS

We thank Nicholas Bhattacharya, Anne Dekas, Bennett Kapili, Michael Kim, Adam Lerer, Hanon McShea, Joshua Meier, Paula Welander, Ellen Zhong, and members of the Peter Kim Laboratory for their helpful comments and discussion. We thank the reviewers for their constructive feedback that improved the manuscript. We thank the Stanford Research Computing Center for providing computational resources and support through the Sherlock cluster. B.L.H. acknowledges the support of the Stanford Science Fellows program. This work was supported by the Chan Zuckerberg Biohub (P.S.K.). A previous version of this manuscript appeared on bioRxiv (https://doi.org/10.1101/2021.06.07.447389).

AUTHOR CONTRIBUTIONS

All authors were involved in project conceptualization and investigation. B.L.H. wrote the software, performed the computational experiments, and wrote the initial paper draft. All authors interpreted the results and wrote the final paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 5, 2021
Revised: November 15, 2021
Accepted: January 12, 2022
Published: February 3, 2022

REFERENCES

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

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44.

Bepler, T., and Berger, B. (2019). Learning protein sequence embeddings using information from structure. In 7th International Conference on Learning Representations, arXiv:1902.08661.

Bepler, T., and Berger, B. (2021). Learning the protein language: evolution, structure, and function. Cell Syst *12*, 654–669.e3.

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat. Biotechnol. *38*, 1408–1414.

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

Chen, H., Zheng, D., Davids, J., Bartee, M.Y., Dai, E., Liu, L., Petrov, L., Macaulay, C., Thoburn, R., Sobel, E., et al. (2011). Viral serpin therapeutics: from concept to clinic. Methods Enzymol *499*, 301–329.

Dawkins, R. (1976). The Selfish Gene (Oxford University Press).

de Visser, J.A.G.M., and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. Nat. Rev. Genet. *15*, 480–490.

Eckert, D.M., and Kim, P.S. (2001). Mechanisms of viral membrane fusion and its inhibition. Annu. Rev. Biochem. *70*, 777–810.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res *47*, D427–D432.

Gong, L.I., Suchard, M.A., and Bloom, J.D. (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. Elife *2*, e00631.

Gould, S.J. (1990). Wonderful Life: The Burgess Shale and the Nature of History (WW Norton & Company).

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Gupta, A., and Adami, C. (2016). Strong selection significantly increases epistatic interactions in the long-term evolution of a protein. PLoS Genet *12*, e1005960.

Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *13*, 845–848.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.

Harrison, S.C. (2008). Viral membrane fusion. Nat. Struct. Mol. Biol. *15*, 690–698.

Hayden, E.J., and Wagner, A. (2012). Environmental change exposes beneficial epistatic interactions in a catalytic RNA. Proc. Biol. Sci. *279*, 3418–3425.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. *32*, 835–845.

Henikoff, S., and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA *89*, 10915–10919.

Hie, B., Peters, J., Nyquist, S.K., Shalek, A.K., Berger, B., and Bryson, B.D. (2020). Computational methods for single-cell RNA sequencing. Annu. Rev. Biomed. Data Sci. *3*, 339–364.

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

Hsu, C., Nisonoff, H., Fannjiang, C., and Listgarten, J. (2022). Learning protein fitness models from evolutionary and assay-labeled data. Nature Biotechnology. https://doi.org/10.1038/s41587-021-01146-5.

Irving, J.A., Steenbakkers, P.J.M., Lesk, A.M., Op den Camp, H.J.M., Pike, R.N., and Whisstock, J.C. (2002). Serpins in prokaryotes. Mol. Biol. Evol. *19*, 1881–1890.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. *8*, 275–282.

Kim, Y., Koehler, F., Moitra, A., Mossel, E., and Ramnarayan, G. (2020). How many subpopulations is too many? Exponential lower bounds for inferring population histories. J. Comp. Biol. *27*, 136–157.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*, 494–498.

Lässig, M., Mustonen, V., and Walczak, A.M. (2017). Predicting evolution. Nat. Ecol. Evol. *1*, 77.

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res *47*, W256–W259.

Livesey, B.J., and Marsh, J.A. (2020). Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol. Syst. Biol. *16*, e9380.

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

Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Jr., Xiong, C., Sun, Z.Z., Socher, R., et al. (2021). Deep neural language modeling enables functional protein generation across families. bioRxiv. https://doi.org/10.1101/2021.07.18.452833.

Maher, C.M., Bartha, I., Weaver, S., di Iulio, J., Ferri, E., Soriaga, L., Lempp, F.A., Hie, B.L., Bryson, B., Berger, B., et al. (2022). Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Science Translational Medicine. https://doi.org/10.1126/scitranslmed.abk3445.

Masuda, N., Porter, M.A., and Lambiotte, R. (2017). Random walks and diffusion on networks. Phys. Rep. *716–717*, 1–58.

Mccandlish, D.M. (2011). Visualizing fitness landscapes. Evolution *65*, 1544–1558.

McInnes, L., and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426.

McLaughlin, P.J., and Dayhoff, M.O. (1973). Eukaryote evolution: a view based on cytochrome *c* sequence data. J. Mol. Evol. *2*, 99–116.

Morris, S.C. (2003). Life's Solution: Inevitable Humans in a Lonely Universe (Cambridge University Press).

Narayan, A., Berger, B., and Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. Nat. Biotechnol. *39*, 765–774.

Piast, M., Kustrzeba-Wójcicka, I., Matusiewicz, M., and Banaś, T. (2005). Molecular evolution of enolase. Acta Biochim. Pol. *52*, 507–513.

Pillai, A.S., Chandler, S.A., Liu, Y., Signore, A.V., Cortez-Romero, C.R., Benesch, J.L.P., Laganowsky, A., Storz, J.F., Hochberg, G.K.A., and Thornton, J.W. (2020). Origin of complexity in haemoglobin evolution. Nature *581*, 480–485.

Potter, S., and Fothergill-Gilmore, L.A. (1993). Molecular evolution: the origin of glycolysis. Biochem. Educ. *21*, 45–48.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with TAPE. Adv. Neural Inf. Process. Syst. *32*, 9686–9698.

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

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

Roberts, T.H., Hejgaard, J., Saunders, N.F.W., Cavicchioli, R., and Curmi, P.M.G. (2004). Serpins in unicellular Eukarya, Archaea, and Bacteria: sequence analysis and evolution. J. Mol. Evol. *59*, 437–447.

Rojas-Pirela, M., Andrade-Alviárez, D., Rojas, V., Kemmerling, U., Cáceres, A.J., Michels, P.A., Concepción, J.L., and Quiñones, W. (2020). Phosphoglycerate kinase: structural aspects and functions, with special emphasis on the enzyme from Kinetoplastea. Open Biol *10*, 200302.

Sharp, P.M., and Hahn, B.H. (2011). Origins of HIV and the AIDS pandemic. Cold Spring Harbor Perspect. Med. *1*, a006841.

Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data—from vision to reality. Euro Surveill *22*, 30494.

Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature *225*, 563–564.

Spence, M.A., Mortimer, M.D., Buckle, A.M., Minh, B.Q., and Jackson, C.J. (2021). A comprehensive phylogenetic analysis of the serpin superfamily. Mol. Biol. Evol. *38*, 2915–2929.

Sutton, T.C. (2018). The pandemic threat of emerging H5 and H7 avian influenza viruses. Viruses *10*, 461.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res *47*, D506–D515.

Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A., et al. (2015). The immune epitope database (IEDB) 3.0. Nucleic Acids Res *43*, D405–D412.

Walensky, R.P., Walke, H.T., and Fauci, A.S. (2021). SARS-CoV-2 variants of concern in the United States-challenges and opportunities. JAMA *325*, 1037–1038.

Wei, C.J., Boyington, J.C., Dai, K., Houser, K.V., Pearce, M.B., Kong, W.P., Yang, Z.Y., Tumpey, T.M., and Nabel, G.J. (2010). Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. Sci. Transl. Med. *2*, 24ra21.

Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. (2016). The physiology and habitat of the last universal common ancestor. Nat. Microbiol. *1*, 16116.

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. *18*, 691–699.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol *19*, 15.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. Sixth International Congress on Genetics *1*, 355–366.

Xu, R., Ekiert, D.C., Krause, J.C., Hai, R., Crowe, J.E., and Wilson, I.A. (2010). Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science *328*, 357–360.

Yu, Y.W., Daniels, N.M., Danko, D.C., and Berger, B. (2015). Entropy-scaling search of massive Biological Data. Cell Syst *1*, 130–140.

Zhang, Y., Aevermann, B.D., Anderson, T.K., Burke, D.F., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C.N., Lee, A.J., et al. (2017). Influenza Research Database: an integrated bioinformatics resource for influenza virus research. Nucleic Acids Res *45*, D466–D474.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Sequences, models, and metadata	This study	https://doi.org/10.5281/zenodo.5590361
UniRef50	(Suzek et al., 2007)	https://www.uniprot.org/help/uniref
Immune Epitope Database	(Vita et al., 2015)	https://www.iedb.org/
NIAID Influenza Research Database	(Zhang et al., 2017)	https://www.fludb.org/
Los Alamos National Laboratory HIV Database	N/A	https://www.hiv.lanl.gov/
UniProt	(UniProt Consortium, 2019)	https://www.uniprot.org/
Software and algorithms		
Code and scripts	This study	https://github.com/brianhie/evolocity and https://doi.org/10.5281/zenodo.5544302
ESM-1b	(Rives et al., 2021)	https://github.com/facebookresearch/esm
TAPE	(Rao et al., 2019)	https://github.com/songlab-cal/tape
Scanpy version 1.6.1	(Wolf et al., 2018)	https://scanpy.readthedocs.io/
scVelo version 0.2.2	(Bergen et al., 2020)	https://scvelo.readthedocs.io/
PhyML version 3.3.20200621	(Guindon et al., 2010)	https://github.com/stephaneguindon/phyml
NumPy version 1.17.2	(Harris et al., 2020)	https://numpy.org/
iTOL	(Letunic and Bork, 2019)	https://itol.embl.de/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Brian L. Hie (brianhie@stanford.edu).

Materials availability

No new materials were generated in this study.

Data and code availability

All data used in our analysis has been deposited to Zenodo and are publicly available as of the date of publication. DOIs are listed in the key resources table. All original code used in our analysis has been deposited to Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table. Our code and links to data are also available on GitHub at https://github.com/brianhie/evolocity. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Deep mutational scan benchmarking

We obtained DMS values, all involving single-residue substitutions, and the corresponding DeepSequence (Riesselman et al., 2018) mutational effect predictions from Livesey and Marsh (2020). To compute mutational effect predictions for ESM-1b and TAPE, we used the evo-velocity score between the wildtype and mutant sequence as described above. As done by Livesey and March, we evaluated the performance of the mutational effect prediction as the absolute value of the Spearman correlation between the algorithm's predicted mutational effect and the value reported by the original DMS study, restricting only to mutants considered by the original DMS studies. We used all DMS studies from Livesey and Marsh for which there were DeepSequence results available. When DMS studies performed multiple types or degrees of selection, we computed the correlation between language model likelihood and observed variant effects for each instance.

Influenza A NP evo-velocity analysis

We obtained 3,304 unique NP sequences from the NIAID Influenza Research Database (https://www.fludb.org) (Zhang et al., 2017). We restricted our analysis to sequences that were sampled from human hosts. Metadata included the year the sequences were sampled and the influenza subtype of the original virus. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously.

We obtained an ordered phylogenetic path from Gong et al. (2013) of H3N2-subtype NP evolution from 1968 to 2007. We computed the ESM-1b evo-velocity score comparing adjacent sequences along this path and plotted the cumulative sum of these scores versus the order in the path (Figure 2D). We also compared the improvement in evo-velocity of this path to that of simulated paths. To simulate paths across our evo-velocity landscape, we began at the same starting sequence, used the same number of steps as the path of Gong et al., and only considered paths that ended in the same cluster of sequences as the end sequence of Gong et al.'s path. We used the transition matrix $\mathbf{Q}$ to define the probability of moving from node to node and we performed 30,000 random walks.

We obtained a phylogenetic tree of all NP sequences considered in the evo-velocity analysis by first aligning sequences with MAFFT followed by approximate maximum-likelihood tree construction using FastTree version 2.1 using a JTT+CAT model. The midpoint-rooted tree was visualized using the iTOL web tool (https://itol.embl.de/) (Letunic and Bork, 2019).

We also projected evo-velocity into one-hot-encoding space to compute a $N|\mathcal{X}|$-dimensional vector $\tilde{\mathbf{v}}_a$ for each sequence as described previously; we then averaged these vectors across all sequences and inspected the top five mutations with the greatest magnitude change in the resulting average. We then located these mutations onto a reference sequence from 1934 H1N1 NP (UniProt ID: P03466), for which linear T-cell epitope data is available through the Immune Epitope Database (https://www.iedb.org/) (Vita et al., 2015). We restricted our consideration to linear epitopes of influenza NP with positive validation in a T-cell assay.

Influenza A HA evo-velocity analysis

We obtained 8,115 unique HA H1 sequences from the NIAID Influenza Research Database (https://www.fludb.org) (Zhang et al., 2017). We restricted our analysis to sequences that were sampled from human hosts. Metadata included the year the sequences were sampled and the influenza subtype of the original virus. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously.

HIV-1 Gag evo-velocity analysis

We obtained 18,018 unique Gag sequences from the LANL HIV sequence database (https://www.hiv.lanl.gov). Metadata included the year the sequences were sampled and the HIV subtype of the original virus. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously. We obtained four SIVcpz Gag sequences with high-quality, manual annotation from UniProt (https://www.uniprot.org/) (UniProt Consortium, 2019). These sequences were obtained from SIVcpz isolates MB66 (UniProt ID: Q1A268), EK505 (UniProt ID: Q1A250), TAN1 (UniProt ID: Q8AII2), and GAB1 (UniProt ID: P17282).

SARS-CoV-2 Spike evo-velocity analysis

We obtained 75,584 unique, full-length Spike sequences from the August 25, 2021 GISAID release (https://www.gisaid.org/) (Shu and McCauley, 2017). Metadata included the date the sequences were sampled. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously. We determined the location of clusters corresponding to known variants-of-concern based on known marker mutations including D614G, N501Y (for B.1.1.7, B.1.351, and P.1), K417N (for B.1.351), P681H (for B.1.1.7), E154K (for B.1.617.1), T478K (for B.1.617.2), and P681R (for B.1.617.2) (Walensky et al., 2021).

Globins evo-velocity analysis

We obtained 6,097 globin sequences from UniProt. We restricted our analysis to eukaryotic sequences within the "globin" family and to sequences between 135 and 155 residues in length, inclusive, which was done based on a clear mode in the distribution of sequence lengths and was meant to preserve mostly homologous sequences in our analysis. Metadata included the taxonomic lineage of each sequence and, for some of the sequences, annotations indicating the type of globin. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously. We obtained the rooted phylogenetic tree of globins and the inferred ancestral sequences from Pillai et al. (2020).

Cytochrome c evo-velocity analysis

We obtained 2,128 cytochrome c sequences from UniProt. We restricted our analysis to eukaryotic sequences within the "cytochrome c" family and to sequences between 100 and 115 residues in length, inclusive, which was done based on a clear mode in the distribution of sequence lengths and was meant to preserve mostly homologous sequences in our analysis. Metadata included the taxonomic lineage of each sequence. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously. We obtained the approximate dates and geologic eons of the emergences of different organisms from Hedges et al. (2015).

Enolase evo-velocity analysis

We obtained 31,901 enolase sequences from UniProt. We restricted our analysis to sequences within the "enolase" family and to sequences between 412 and 448 residues in length, inclusive, which was done based on a clear mode in the distribution of sequence lengths and was meant to preserve mostly homologous sequences in our analysis. Metadata included the taxonomic lineage of each sequence. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously.

We obtained unrooted phylogenetic trees of enolase based on the subset of our UniProt sequences with high-quality, manual annotation. We then performed a multiple sequence alignment with MAFFT and performed phylogenetic reconstruction on the alignment with PhyML version 3.3.20200621 using a JTT model with gamma-distributed among-site rate variation and empirical state frequencies (Guindon et al., 2010). The unrooted tree was visualized using the iTOL web tool.

PGK evo-velocity analysis

We obtained 30,455 PGK sequences from UniProt. We restricted our analysis to sequences within the "phosphoglycerate kinase" family and to sequences between 385 and 420 residues in length, inclusive, which was done based on a clear mode in the distribution of sequence lengths and was meant to preserve mostly homologous sequences in our analysis. Metadata included the taxonomic lineage of each sequence. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously.

We obtained unrooted phylogenetic trees of enolase based on the subset of our UniProt sequences with high-quality, manual annotation. We then performed a multiple sequence alignment with MAFFT and performed phylogenetic reconstruction on the alignment with PhyML using a JTT model with gamma-distributed among-site rate variation and empirical state frequencies. The unrooted tree was visualized using the iTOL web tool.

Serpins evo-velocity analysis

We obtained 22,737 serpin sequences from UniProt. We restricted our analysis to sequences within the "serpin" family and to sequences between 300 and 525 residues in length, inclusive, which was done based on a clear mode in the distribution of sequence lengths and was meant to preserve mostly homologous sequences in our analysis. Metadata included the taxonomic lineage of each sequence. We performed KNN graph construction, evo-velocity computation, root prediction, diffusion pseudotime estimation, and UMAP velocity projection as described previously.

We obtained unrooted phylogenetic trees of enolase based on the subset of our UniProt sequences with high-quality, manual annotation. We then performed a multiple sequence alignment with MAFFT and performed phylogenetic reconstruction on the alignment with PhyML using a JTT model with gamma-distributed among-site rate variation and empirical state frequencies. The unrooted tree was visualized using the iTOL web tool.

Evo-velocity benchmarking metrics

We performed control experiments to test how modifying different components of our evo-velocity implementation affected the ability of diffusion pseudotime to predict the directionality of evolution. For proteins with largely continuous evolution and with known sampling date (i.e., influenza A NP, influenza A HA, and SARS-CoV-2 Spike), we computed the Spearman correlation between sampling date and pseudotime. For cytochrome c, we tested the ability to predict taxonomic order, as determined by location in the fossil record (Figure 4F), by computing the Spearman correlation between pseudotime and the ordinal value of each taxonomic class. For the remaining proteins, we tested the ability for pseudotime to separate categorical taxonomic classes according to their known evolutionary order; to quantify this separation, we use the area under the receiver-operating characteristic curve (AUROC) where we assign 0 to the first class and 1 to the second class. For HIV-1 Gag, we used the combined subtypes B and C as the first class and circulating recombinant form BC as the second class. For globins, we used neuroglobin as the first class and Hbβ as the second class. For serpins, we used prokaryota as the first class and eukaryota as the second class. For enolase and PGK, we used archaea as the first class and eukaryota as the second class.

UniRef50 sequence similarity computational control

We wanted to quantify if our evo-velocity results, including evo-velocity pseudotime, could be explained by sequence similarity to the training set. We obtained this training set from ftp://ftp.uniprot.org/pub/databases/uniprot/previous_releases/release-2018_03/uniref/. We identified representative sequences in UniRef50 by searching for the literal presence of the sequence within UniRef50 or by mapping the protein accession information to UniProt IDs, if available, and then mapping the UniProt IDs to the corresponding UniRef50 cluster representative. Then, for each sequence in our evo-velocity analysis, we computed the sequence similarity score to each representative sequence in UniRef50 and took the maximum of these scores. To compute the sequence similarity score, we used the similarity ratio implemented by the fuzzywuzzy Python package version 0.18.0, which is based on the Levenshtein distance between two sequences and is normalized to take values between 0% and 100%, inclusive.

To perform the control experiment, we filtered out sequences with 80% or less sequence similarity to the training set, thereby excluding sequences that are far from the sequences considered by ESM-1b. We then evaluated the Spearman correlation between the similarity scores and pseudotime, both in terms of the directionality of the correlation (e.g., a positive correlation indicates that similarity to UniRef50 could be explaining pseudotime) and also in terms of the change in this correlation compared to the correlation

obtained on the full set of sequences (Table S3). We also evaluated the ability for the overall pseudotemporal patterns, quantified using the benchmarking metrics described above, to reproduce those found when analyzing the full set of sequences.

TAPE reproducibility computational control

To see how robust our evo-velocity results were to the language model used to estimate the mutational likelihoods, we obtained the TAPE transformer model as described above. We performed the evo-velocity analysis by keeping the KNN graph structure the same as in the ESM-1b analysis but using the evo-velocity scores obtained by the TAPE likelihoods. All other downstream analyses were also kept the same.

Non-epistatic substitution matrix computational control

We also tested velocity models based on amino-acid substitution matrices, which compute scores without any sequence context or epistatic information. We obtained three substitution matrices: BLOSUM62 (Henikoff and Henikoff, 1992), JTT (Jones et al., 1992) and WAG (Whelan and Goldman, 2001). Rather than weight edges in the KNN network based on language model pseudolikelihood scores, we instead use a score based on the mean of the substitution matrix-defined mutation scores; more specifically, given a substitution matrix that outputs a scalar for a given amino acid pair, which we denote as $s : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, we compute velocity scores as $v'_{ab} \underline{\underline{\text{def}}} \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} s\left(x_i^{(a)}, x_i^{(b)}\right)$. All three substitution matrices we consider are symmetric, so $v'_{ab} = v'_{ba}$ (resulting in an undirected, weighted graph). The rest of the analysis, including softmax-based transition probability computation and subsequent diffusion analysis, remained the same.

Negative control experiments

We performed additional controls meant to assess the value of incorporating evolutionary information into the evo-velocity algorithm. We also used the same benchmarking metrics described above. To assess the value of a KNN graph based on Euclidean distance in language model-embedding space, we instead constructed the KNN graph using Hamming distance in one-hot-encoded embedding space, using the same nearest-neighbors implementation provided by Scanpy; we used the same value of k and the same downstream steps as the original analysis.

To test the value of analysis based on a structured embedding space, we instead treated all pairs of sequences as having unit distance between them; to construct the KNN graph (while preserving the computationally tractability of downstream analysis), we broke ties by randomly selecting a set of edges using the same value of k, followed by the same downstream steps as the original analysis. Pseudotime values for this control experiment were averaged for each sequence across three random seeds and subsequently rescaled.

To test the value of ESM-1b pretraining on UniRef50, we instead computed velocity scores using ESM-1b with randomly initialized weights, followed by all steps taken in the original analysis. Pseudotime values for this control experiment were averaged for each sequence across three random initializations of ESM-1b and subsequently rescaled.

To test the value of biasing network diffusion analysis with evolutionary information, we instead assigned unit velocity to all edges in the network, i.e., $v_{ab} \underline{\text{def}} 1$, followed by all steps taken in the original analysis.

Downsampling benchmark experiments

We tested the robustness of evo-velocity to downsampling of the original sequence landscape. We downsampled to 75%, 50%, 25%, or 10% of the original landscape without replacement (the number of samples in each resulting landscape is provided in Data S2), repeating across three random seeds. Our first experiment sampled sequences with uniform probability across the range described above. Our second experiment preferentially weighted sequences with the goal of preserving those that are more recent in evolutionary time, simulating the higher probability of missing sequences that are early in evolution (though we note that the original landscapes are most likely already biased to newer sequences). For the landscapes of influenza A NP, influenza A HA, and SARS-CoV-2 Spike, we weighted sequences by their rank order according to known sampling time, so that later sequences have a higher weight. For the HIV-1 Gag landscape, we placed unit weight on primary subtypes and double this weight on circulating recombinant forms. For the globin landscape, we placed unit weight on all globin types except the hemoglobin subunits, on which we placed double the unit weight. For the cytochrome c landscape, we placed unit weight on other eukaryota, double the unit weight on viridiplantae and fungi, and triple the unit weight on arthropoda, chordata, and mammalia. For the landscapes of serpins, enolase, and PGK, we placed unit weight on archaea, double the unit weight on bacteria, and triple the unit weight on eukaryota. Once weights were placed on sequences, they were scaled to valid probabilities by normalized the weights to sum to unity across all sequences. We used the weighted random sampling functionality provided by the NumPy Python package version 1.17.2.

QUANTIFICATION AND STATISTICAL ANALYSIS

Language models

In this paper, we implement evo-velocity with masked language models, which are trained by masking certain residues in the input and predicting these residues in the output. For a sequence $\mathbf{x} \in \mathcal{X}^N$, where $\mathcal{X}$ is the set of amino acids and N is the sequence length, the masked language modeling objective implicitly models a distribution over sequences through conditional likelihoods

$p(x_i|\mathbf{x}_{[N]\ \{i\}})$ where $\mathbf{x}_{[N]\ \{i\}}$ denotes the sequence without the residue at position i, sometimes referred to as the sequence context. Typically, these language models also learn a latent variable $\mathbf{z}_i \in \mathbb{R}^D$ by learning a function $f : \mathcal{X}^{N-1} \rightarrow \mathbb{R}^D$ where $\mathbf{z}_i \underline{\text{def}}\ f(\mathbf{x}_{[N]\ \{i\}})$ such that $p(x_i|\mathbf{x}_{[N]\ \{i\}}, \mathbf{z}_i) = p(x_i|\mathbf{z}_i)$.

We use two large-scale language models trained with a masked objective. We used the ESM-1b model (Rives et al., 2021) (obtained from https://github.com/facebookresearch/esm) trained on the March 2018 release of UniRef50 (Suzek et al., 2007). We also used the TAPE transformer model (Rao et al., 2019) (obtained from https://github.com/songlab-cal/tape) trained on the Pfam database release 32.0 (El-Gebali et al., 2019). Unless otherwise stated, we used ESM-1b as the default model for our experiments.

Evo-velocity score computation

We compute an evo-velocity score that compares two sequences $\mathbf{x}^{(a)}$ and $\mathbf{x}^{(b)}$ as

$$v_{ab}\ \underline{\text{def}}\ \frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}}\left[\log p\left(x_i^{(b)}\middle|\mathbf{z}_i^{(a)}\right) - \log p\left(x_i^{(a)}\middle|\mathbf{z}_i^{(b)}\right)\right],$$

where $\mathcal{M}\underline{\text{def}}\{i\ : x_i^{(a)} \neq x_i^{(b)}\}$ is the set of positions at which the amino acid residues disagree. We designed the evo-velocity score based on masked-language-model pseudolikelihoods (Hsu et al., 2022) to efficiently approximate the change in likelihood of mutating sequence $\mathbf{x}^{(a)}$ to $\mathbf{x}^{(b)}$ and vice versa. The evo-velocity score is positive if moving from $\mathbf{x}^{(a)}$ to $\mathbf{x}^{(b)}$ is more favorable, negative if moving from $\mathbf{x}^{(b)}$ to $\mathbf{x}^{(a)}$ is more favorable (so that $v_{ab} = -v_{ba}$), and zero if they are equal.

In practice, $\mathbf{x}^{(a)}$ and $\mathbf{x}^{(b)}$ can disagree in length, so we first perform a global pairwise sequence alignment using the pairwise2 module in the Biopython Python package version 1.76 with a uniform substitution matrix and alignment parameters meant to discourage the introduction of sequence gaps (following the Biopython recommendations, we use a match score of 5, a mismatch penalty of -4, a gap-open penalty of -4, and a gap-extension penalty of -0.1). We ignore positions involving alignment gaps when computing the evo-velocity score, i.e., the evo-velocity score is only based on substitutions, since modeling the effect of an insertion or a deletion is less well defined when using a masked language model to predict mutations. We do not include gap characters when computing language model likelihoods.

Constructing the sequence similarity network and evo-velocity transition matrix

To construct the sequence similarity network, we first use the language model to obtain a sequence embedding $\mathbf{z}^{(a)}\ \underline{\text{def}}\ \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i^{(a)}$ for each sequence $\mathbf{x}^{(a)}$ in the set of sequences-of-interest (for example, proteins within the same family) of size M. We use ESM-1b to compute the embeddings for each sequence as the 1,280-dimensional output of the last (i.e., the 33rd) hidden layer of the language model.

We then construct a directed graph where each node corresponds to a sequence and we connect a node to its k-nearest neighbors based on the Euclidean distance in the language model embedding space in $\mathbb{R}^D$. We can then use the evo-velocity scores and the KNN graph to construct a transition matrix $\mathbf{Q} \in \mathbb{R}^{M\times M}$, where

$$q_{ab}\ \underline{\text{def}}\ \frac{\exp(v_{ab})}{\sum_{b'\in\mathcal{N}(\mathbf{x}_a)}\exp(v_{ab'})}$$

is the entry in the ath row and bth column of $\mathbf{Q}$ and $\mathcal{N}(\bullet)$ denotes the set of the neighbors in the KNN graph. Note that $\sum_{b\in[M]} q_{ab} = 1$.

In all our experiments, we use the embedding function learned by the ESM-1b language model. To construct the KNN graph, we use the functionality provided by the Scanpy Python package version 1.6.1 (Wolf et al., 2018). In practice, higher values of k result in smoother, less noisy landscapes at the cost of higher computational effort. We find that values of k around 30 to 50 (our package defaults to 50) provide a good balance between robustness to noise and computational efficiency (though analyses involving less sequences overall or more homogeneous sequences can also tolerate lower values of k to speed up analysis); the 30-50 range has also shown good empirical performance in other KNN-based analyses that require robust estimation of the biological landscape (Narayan et al., 2021). In this paper, we use the values $k = 30$ for our cytochrome c and Spike experiments, $k = 40$ for our NP and Gag experiments, and $k = 50$ for our HA, globin, enolase, PGK, and serpin experiments. In general, we find that lower values of k are sufficient for densely-sampled landscapes or shorter timescales, whereas larger values of k are required for more sparsely-sampled landscapes over longer timescales.

Network diffusion analysis and predicting roots

To find the root nodes, we can use the fixed points of a diffusion process based on the transition matrix $\mathbf{Q}$ (Bergen et al., 2020; Masuda et al., 2017). Given a diffusion probability vector $\boldsymbol{\mu}^{(t)}$, we can find roots by running a diffusion process until a fixed point, i.e., $\boldsymbol{\mu}^{(\infty)} = \mathbf{Q}^{\mathsf{T}}\boldsymbol{\mu}^{(\infty)}$ (note that we take the transpose of the transition matrix to "reverse" the diffusion process, since our goal is to find the root nodes). We take the highest values of $\boldsymbol{\mu}^{(\infty)}$ to identify the root nodes, where we obtain $\boldsymbol{\mu}^{(\infty)}$ as the eigenvector of $\mathbf{Q}^{\mathsf{T}}$ corresponding to an eigenvalue of 1. By default, we use a cutoff at the 98$^{\text{th}}$ percentile of values in $\boldsymbol{\mu}^{(\infty)}$ to define the set of root nodes, as has been done previously (Bergen et al., 2020). We assume $\mathbf{Q}$ corresponds to a strongly connected directed graph, which is true if the KNN graph consists of a single connected component (and which was true for all of our analyses), since each undirected edge in the original KNN graph leads to two directed edges in the velocity graph; if the graph is strongly connected, then there is a unique value of

$\boldsymbol{\mu}^{(\infty)}$ (Masuda et al., 2017). We scale the final values of the diffusion vector $\boldsymbol{\mu}^{(\infty)}$ to take values between 0 and 1, inclusive, and use the diffusion-based root estimation procedure implemented by the scVelo Python package version 0.2.2 (Bergen et al., 2020).

Diffusion pseudotime computation

We use diffusion pseudotime (DPT) to order sequences in evolutionary time. DPT is described in detail by Haghverdi et al. (2016) and is closely related to the geodesic distance between two nodes in a graph. As done by Haghverdi et al., we denote the DPT score between a root node $\mathbf{x}^{(root)}$ and a node $\mathbf{x}$ as $dpt(\mathbf{x}^{(root)}, \mathbf{x})$, which takes scaled values between 0 and 1, inclusive. We use the graph encoded by the transition matrix $\mathbf{Q}$. Since the root-prediction analysis described above can yield potentially multiple roots, we define evo-velocity pseudotime as the average of DPT scores across the set of all root nodes $\mathcal{R}$, i.e.,

$$pseudotime(\mathbf{x}) \underline{\underline{\text{def}}} \frac{1}{|\mathcal{R}|} \sum_{\mathbf{x}^{(root)} \in \mathcal{R}} dpt\left(\mathbf{x}^{(root)}, \mathbf{x}\right).$$

We use the DPT implementation provided by the Scanpy Python package.

Plotting, data visualization, and statistical analysis

We used the UMAP algorithm (McInnes and Healy, 2018) to visualize the KNN graph in two dimensions. All UMAP visualizations were obtained using the umap-learn Python package version 0.4.6 as wrapped by Scanpy. We generated boxplots using the seaborn Python package version 0.11.1; in all of our boxplots, the box extends from the first to third quartile, a horizontal line is drawn at the median, and whiskers extend to 1.5 times the interquartile range. We used the scipy version 1.4.1 Python package to compute correlations and statistical tests. A P value of less than 1×10^{-308} indicates a value that was below the floating-point precision of our computer.

Embedding transfer

We can project evo-velocity, as encoded by the transition matrix $\mathbf{Q}$, into an arbitrary embedding space (assuming that embeddings are available for all sequences) as done previously (Bergen et al., 2020). For a sequence $\mathbf{x}^{(a)}$ and $\mathbf{x}^{(b)}$, we denote the respective embeddings as $\mathbf{z}^{(a)}$ and $\mathbf{z}^{(b)}$. We then first compute the cosine-normalized translation vector separating sequences connected in the KNN graph, i.e.,

$$\boldsymbol{\delta}_{ab} \underline{\underline{\text{def}}} \frac{\mathbf{z}_b - \mathbf{z}_a}{\|\mathbf{z}_b - \mathbf{z}_a\|_2}$$

and we obtain the velocity projections as the expected displacement with respect to $\mathbf{Q}$, i.e.,

$$\tilde{\mathbf{v}}_a \underline{\underline{\text{def}}} \sum_{b \neq a} \left(q_{ab} - \frac{1}{M}\right) \boldsymbol{\delta}_{ab}.$$

We use two main interpretable embedding spaces in our downstream analysis. The first is two-dimensional UMAP space, in which evo-velocity can be visualized as two-dimensional vectors. Once these vectors are computed, we use the streamplot and quiver plot functionality of the matplotlib Python package version 3.3.3 to visualize evo-velocity. The second interpretable embedding space we consider is one-hot-encoded sequence space, which we use to identify mutations that are associated with large changes in evo-velocity. To project evo-velocity into sequence space, we first construct a multiple sequence alignment of all M sequences using MAFFT version 7.475. A sequence $\mathbf{x}$ is then embedded into a one-hot-encoded vector $\tilde{\mathbf{z}} \in \{0,1\}^{\tilde{N}|\mathcal{X}|}$, where $\tilde{N}$ is the length of the alignment. The velocity projections take values in $\mathbb{R}^{\tilde{N}|\mathcal{X}|}$, where we interpret each dimension as corresponding to a given residue in $\mathcal{X}$ at a given site in $[\tilde{N}]$.