A Structured Observation Distribution for Generative Biological Sequence Prediction and Forecasting

Eli N. Weinstein [1] **Debora S. Marks** [2 3]

Abstract

Generative probabilistic modeling of biological sequences has widespread existing and potential application across biology and biomedicine, from evolutionary biology to epidemiology to protein design. Many standard sequence analysis methods preprocess data using a multiple sequence alignment (MSA) algorithm, one of the most widely used computational methods in all of science (Van Noorden et al., 2014). However, as we show in this article, training generative probabilistic models with MSA preprocessing leads to statistical pathologies in the context of sequence prediction and forecasting. To address these problems, we propose a principled drop-in alternative to MSA preprocessing in the form of a structured observation distribution (the "MuE" distribution). We prove theoretically that the MuE distribution comprehensively generalizes popular methods for inferring biological sequence alignments, and provide a precise characterization of how such biological models have differed from natural language latent alignment models. We show empirically that models that use the MuE as an observation distribution outperform comparable methods across a variety of datasets, and apply MuE models to a novel problem for generative probabilistic sequence models: forecasting pathogen evolution.

1. Introduction

High-throughput sequencing is pervasive across biology and biomedicine, and critical to both past and ongoing discoveries and technological advancements. Analyzing large scale sequence data, making predictions about unobserved or future sequences, and generating new functional sequences, are major and growing challenges with relevance to epidemiology (predicting pathogen evolution), immunology (characterizing antibody repertoires), molecular evolution (mapping substructure within protein families), protein design, and many more subfields of biology and biomedicine. In principal, generative probabilistic modeling enables (a) modular and uncertainty-aware data analysis, (b) formal mathematical statement of underlying assumptions, and (c) generation of new samples, which in the case of sequences can be synthesized and tested in the laboratory (taking advantage of recent rapid progress in high-throughput synthesis) (Kucukelbir et al., 2017; Russ et al., 2020). However, although machine learning and statistics offer an extraordinary array of generative probabilistic models, extending existing methods to apply to biological sequences while accounting for domain-specific prior knowledge is nontrivial.

When analyzing biological sequence data, a standard approach is to preprocess the data before building any models by constructing a multiple sequence alignment (MSA). MSA algorithms are among the most widely used methods in all of science; according to a 2014 analysis, the 10th most cited scientific article of all time is an MSA algorithm, ahead of all other computational data analysis and statistics articles (Van Noorden et al., 2014; Thompson et al., 1994; 1997). Recent major advances in machine learning and statistical methods for protein structure prediction, variant effect prediction for clinical genetics, protein design, epidemiological tracking, and more have continued to rely on MSAs (Marks et al., 2011; Frazer et al., 2020; Russ et al., 2020; Hadfield et al., 2018). Although MSAs are a powerful tool for understanding sequence evolution, in Section 4.1 of this article we show that employing MSAs as preprocessing introduces statistical pathologies in the context of generative sequence prediction and forecasting.

As a principled, drop-in alternative to MSA preprocessing, this article provides a structured observation distribution for biological sequences, the "mutational emission" ("MuE") distribution. Observation distributions are a common general-purpose technique for extending continuous-space models to other types of data, perhaps most familiar

[1]Program in Biophysics, Harvard University, Cambridge, MA, USA [2]Department of Systems Biology, Harvard Medical School, Boston, MA, USA [3]Broad Institute of Harvard and MIT, Cambridge, MA, USA. Correspondence to: Eli N. Weinstein <eweinstein@g.harvard.edu>, Debora S. Marks <debbie@hms.harvard.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

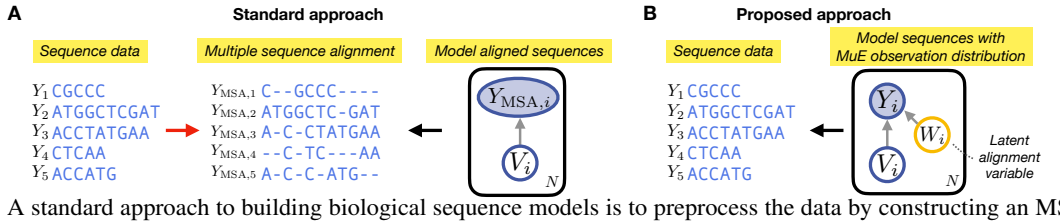

Figure 1. (A) A standard approach to building biological sequence models is to preprocess the data by constructing an MSA. (B) We propose modifying the model instead of the data using the MuE distribution.

in the context of generalized linear models, where they are sometimes also referred to as "error", "emission", or "output" distributions. For instance, to predict count data, one might use a Poisson as an observation distribution, or to predict positive continuous data, one might use a Gamma. Good observation distributions account for both the support of the data and common forms of variability or noise in the data. For biological sequences, we propose using the MuE as an observation distribution. The MuE takes the form of a latent alignment model in which the regressor sequence can also be latent (Deng et al., 2018).[1]

The major contributions of the article are (1) identification of statistical pathologies introduced by widely-used MSA preprocessing methods, (2) a drop-in general purpose alternative, the MuE distribution, (3) a unified and comprehensive theoretical framework for cataloging and rederiving existing biological latent alignment models from the MuE and (4) a novel application of generative probabilistic sequence models enabled by these advancements: forecasting pathogen evolution. At the most practical level, our approach provides a complete recipe for applying one's generative model of choice to biological sequence data while avoiding the pathologies of MSA preprocessing: add a MuE.

2. Method

2.1. Background: MSA Preprocessing

MSA algorithms are applied to families of evolutionarily related biological sequences (proteins, RNA or DNA) in order to infer sites in each sequence that are likely to be related to one another, meaning that they descend from a common ancestor. MSAs can be used as the basis for extrapolation: for instance, knowledge about one region in one sequence can be used to make guesses about related regions in related sequences. MSAs can also be used to understand biological function: for instance, if particular amino acids at particular sites are highly conserved across sequences, it may be evidence that they are crucial to biological function. Generative probabilistic models of MSAs have seen widespread success on these and many other tasks, including predicting the clinical impacts of genetic mutations, inferring three-dimensional protein and RNA structure, and designing new proteins (Frazer et al., 2020; Marks et al., 2011; Weinreb et al., 2016; Russ et al., 2020). We next briefly describe how such MSA-based models are built, as well as their advantages and flaws. In Section 2.2 we introduce our alternative, MuE observation models, which directly generate sequences rather than MSAs. MuE observation models infer related sites but also simultaneously (1) account for uncertainty in which sites are related, (2) allow rigorous model evaluation and (3) enable prediction and forecasting of sequences.

Let $\{Y_1, \ldots, Y_N\}$ be a dataset of N sequences, which may each be different in length, and let $\mathcal{B}$ denote the alphabet (e.g. $\mathcal{B} = \{A, T, G, C\}$ for DNA). MSA algorithms convert the sequence dataset into an N by J matrix, an MSA, adding gap symbols "$-$" such that sites in the same matrix column are those inferred to be related (Figure 1A). Mathematically, MSA algorithms can be summarized as nonlinear functions f_{MSA} that take in datasets of sequences and return processed datasets, $\{Y_{\text{MSA},1}, \ldots Y_{\text{MSA},N}\} := f_{\text{MSA}}(\{Y_1, \ldots, Y_N\})$; for each $i \in \{1, \ldots, N\}$, we have $Y_{\text{MSA},i} \in (\mathcal{B} \cup \{-\})^J$. Note J itself will depend on the input dataset.

Preprocessing sequence data by constructing an MSA is useful in that it (1) converts the data into a matrix, and (2) adjusts for common sources of variability in biological sequence data, in particular insertion and deletion mutations. MSA preprocessing makes building statistical models of sequences more straightforward. For instance, starting from an arbitrary model p_θ that generates continuous matrices $V_i \in \mathbb{R}^{J \times (B+1)}$, where $B := |\mathcal{B}|$, one general strategy is to employ a softmax linker function and a categorical observation distribution ($\text{softmax}(V_i)_j := \exp(V_{i,j,b}) / \sum_{b'} \exp(V_{i,j,b'})$ for $j \in \{1, \ldots, J\}$). The complete approach is (Figure 1A),

$$\begin{aligned} &\text{Preprocess: } \{Y_{\text{MSA},1}, \ldots Y_{\text{MSA},N}\} := f_{\text{MSA}}(\{Y_1, \ldots, Y_N\}), \\ &\text{Model: } V_i \sim p_\theta \\ &Y_{\text{MSA},i} \sim \text{Categorical}(X_i := \text{softmax}(V_i)). \end{aligned} \tag{1}$$

By allowing arbitrary p_θ, this method enables, for example, the application of generative image models (such as variational autoencoders) to biological sequence data (Riesselman et al., 2018). However, as we describe in depth in Section 4.1, MSA preprocessing introduces substantial

[1] We will refer to biological alignments (diagrammatic representations of relatedness between sequences) as "multiple sequence alignments" (Durbin et al., 1998). We will refer to machine learning alignments (latent variables which indicate which positions in one sequence generate which positions in another sequence) as "latent alignments" (Deng et al., 2018).

problems: each row of the output matrix $Y_{\text{MSA},i}$ depends via f_{MSA} on the entire input dataset $\{Y_1, \ldots, Y_N\}$ and we cannot know ahead of time how future raw data Y_{N+1} will change preprocessed past data $Y_{\text{MSA},i \leq N}$. This makes likelihood-based model evaluation on newly observed or heldout data ill-defined.

2.2. The Mutational Emission Distribution

As a drop-in alternative to MSA preprocessing, we introduce the "mutational emission" ("MuE") distribution. The MuE can be used in place of the Categorical observation distribution in Equation 1,

$$\begin{aligned} &\text{Model: } V_i \sim p_\theta \\ &Y_i \sim \text{MuE}(X_i := \text{softmax}(V_i), c, \ell, a^{(0)}, a^{(t)}), \end{aligned} \tag{2}$$

where c, ℓ, $a^{(0)}$, and $a^{(t)}$ are parameters of the MuE, and $V_i \in \mathbb{R}^{M \times D}$ where M and D are hyperparameters rather than dimensions of the input data. The MuE avoids the pathologies of MSA preprocessing by directly generating complete, variable-length sequences (Figure 1B). We refer generically to models that use a MuE observation distribution, such as Equation 2, as "MuE observation" models. (See Figure S1 for a diagram of MuE observation models and Table S1 for a notation reference.) In the limiting case where X_i is a one-hot encoding of a sequence (i.e. $X_{i,m,d} \in \{0, 1\}$ and $\sum_d X_{i,m,d} = 1$), the MuE can be interpreted biologically as generating a mutant Y_i of the "ancestral" sequence X_i, with some probability of insertion and deletion mutations (controlled by c, $a^{(0)}$, and $a^{(t)}$) and of substitution mutations (controlled by ℓ) (Section 2.3). A latent variable W_i in the MuE determines which positions in the regressor X_i – intuitively, which sites in the "ancestral" sequence – generate which positions in Y_i, and can be interpreted as defining a pairwise alignment between X_i and Y_i. The latent variables $W_1, \ldots, W_N$ define a multiple sequence alignment of the dataset $Y_1, \ldots, Y_N$ (Section 4.2). Intuitively, the MuE "adds in", through a generative process, the same mutations that MSA algorithms are intended to "filter out" of the data via preprocessing.

The MuE is a hidden Markov model (HMM) with block-structured emission and transition matrices. Let Δ_D denote the $D-1$ dimensional simplex, $\Delta_D := \{v : v \in \mathbb{R}^D, v_d \geq 0, \sum_{d=1}^D v_d = 1\}$.

Definition 2.1 (MuE) $\text{MuE}(x, c, \ell, a^{(0)}, a^{(t)})$ *is an HMM with $K = 2M + 1$ latent states. The initial probability of each latent state is given by $a^{(0)} \in \Delta_K$, the latent state transition matrix is $a^{(t)} \in (\Delta_K)^K$, and the emission matrix is $\tilde{x} \in (\Delta_D)^K$. The matrices have block structure*

$$\tilde{x} := \begin{bmatrix} x \\ c \end{bmatrix} \cdot \ell, \quad a^{(t)} := \begin{bmatrix} A^{(1,1)} & A^{(1,2)} \\ A^{(2,1)} & A^{(2,2)} \end{bmatrix},$$

where $x \in (\Delta_D)^M$, $c \in (\Delta_D)^{M+1}$, $\ell \in (\Delta_B)^D$, $A^{(1,1)} \in \mathbb{R}^{M \times M}$, and $A^{(2,2)} \in \mathbb{R}^{(M+1) \times (M+1)}$. The transition matrix must satisfy Condition 2.2.

Condition 2.2 (Biological latent alignments) *Entries of $A^{(1,1)}, A^{(1,2)}, A^{(2,1)}$ and $A^{(2,2)}$ below the main diagonal must be zero. Entries of $A^{(1,1)}$ and $A^{(1,2)}$ on the main diagonal must also be zero.*

Condition 2.2, an upper triangular restriction, is illustrated in Figure 2A and justified in depth in Section 4.2. We use w to denote a latent state path taken by the HMM, while W_i denotes the specific latent state path taken when generating sequence Y_i given X_i following $Y_i \sim \text{MuE}(X_i, c, \ell, a^{(0)}, a^{(t)})$.

2.3. Biological Interpretation of the MuE

To describe the biological interpretation of the MuE and its parameters, we consider examples of different latent paths $w = (w_1, \ldots, w_L)$ through state space and the sequences $Y \sim p_{\text{MuE}}(y|x, w)$ that these paths will generate (Figure 2B). Assume to start that $D = B$ and $\ell = I_B$, where I_B is the $B \times B$ identity matrix, and consider the limiting case where x is a one-hot encoding of a sequence (in Figure 2B, the DNA sequence `TACGC`). We consider three example w values:

1. $w = (1, 2, \ldots, M)$ (no mutation). The generated Y will be an exact copy of x, i.e. $Y = x$ if Y is represented as a one-hot encoding (Figure 2B top).
2. $w = (1, \ldots, m-1, m+1, \ldots, M)$ (deletion). The generated Y will be missing the mth letter of x, i.e. $Y = (x_1, \ldots, x_{m-1}, x_{m+1}, \ldots, x_M)$ (Figure 2B middle).
3. $w = (1, \ldots, m, M+m+1, m+1, \ldots, M)$ (insertion). The generated Y will have an additional letter inserted after the mth letter of x, with a probability over letters determined by c_{m+1}, i.e. $Y = (x_1, \ldots, x_m, S, x_{m+1}, \ldots, x_M)$ where $S \sim \text{Categorical}(c_{m+1})$ (Figure 2B bottom).

Condition 2.2 guarantees that the states $k \in \{1, \ldots, M\}$ corresponding to x are each visited at most once and in sequential order. Paths such as $\{1, \ldots, m, m, \ldots, M\}$ (repeat) and $\{1, \ldots, m+1, m, \ldots, M\}$ (backtracking) are not allowed under Condition 2.2. More general matrices $\ell \in (\Delta_B)^D$ allow for substitution mutations, with the probability of converting from letter d to letter b given by $\ell_{d,b}$. For example, if $w = (1, \ldots, M)$, then $Y \sim \text{Categorical}(x \cdot \ell)$, that is Y is a mutant of x with substitution probabilities determined by ℓ and no insertion or deletion mutations.

MuE observation models directly generalize models that use MSA preprocessing in the special case where the dataset

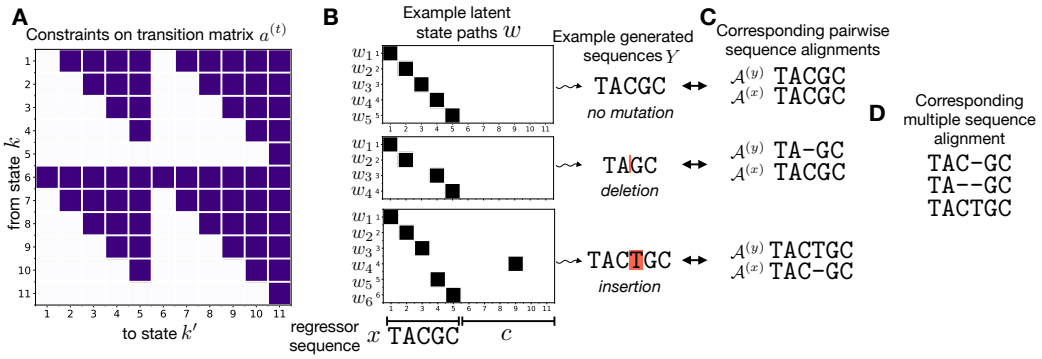

Figure 2. (A) Condition 2.2 allows only the positions of $a^{(t)}$ in dark purple to be non-zero. (B) Example latent state paths w taken by the Markov model in the MuE, and sequences Y that they can generate, given x is a one-hot encoding of the DNA sequence TACGC. Rows correspond to positions $1, \ldots, L$, columns correspond to latent states $1, \ldots, K$. (C) w defines a pairwise alignment between X and Y via Definition 4.3. (D) The collection of w values describe a multiple sequence alignment of the generated sequences Y (Section 4.2).

sequences are all the same length and the MSA algorithm does not add any gap symbols (that is, when $f_{\text{MSA}}(\cdot)$ is the identity). Assume $D = B$, and consider the "no mutation limit" where $\ell = I_B$, $a_1^{(0)} = 1$, and $A_{m,m+1}^{(1,1)} = 1$ for all $m \in \{1, \ldots, M-1\}$. In this case we find, for samples Y of length M, that $Y \sim \text{MuE}(x, c, \ell, a^{(0)}, a^{(t)})$ simplifies to $Y \sim \text{Categorical}(x)$. Thus Equation 2 and Equation 1 become equivalent. In practice, we typically select priors on the MuE to favor the no mutation limit, since it serves as a null hypothesis.

2.4. Inference

The marginal likelihood of the MuE with the latent state variable of the HMM integrated out, $p_{\text{MuE}}(y|x, c, \ell, a^{(0)}, a^{(t)})$, is analytically tractable via the HMM forward algorithm and differentiable. The standard forward algorithm requires $\mathcal{O}(L)$ sequential matrix multiplications, where L is the length of the sequence (typically a few hundred amino acids in our setting), but it can also be parallelized to achieve $\mathcal{O}(\log L)$ time (Särkkä & García-Fernández, 2020; Rush, 2020). Using the MuE marginal likelihood allows inference with automatic differentiation variational inference, stochastic gradient MCMC, and related scalable approximate Bayesian inference algorithms (Section S4.1) (Kucukelbir et al., 2017; Welling & Teh, 2011). We have made available an implementation of the MuE distribution as part of the probabilistic programming language Pyro, making it straightforward to explore different MuE observation models and inference methods (`https://docs.pyro.ai/en/dev/contrib.mue.html`, Section S4.2) (Bingham et al., 2019).

3. Related Work

Methods that use MSA preprocessing. MSA preprocessing is widely used as a starting point for biological sequence data analysis, perhaps most commonly in combination with other non-probabilistic analysis methods. One very common class of probabilistic methods that nearly always use MSA preprocessing is phylogenetic models, which are central to evolutionary biology and genomic epidemiology, and widely used in nearly every other area of biology (Hadfield et al., 2018; Felsenstein, 2004). Another is fitness models, including Potts models and variational autoencoder models, which are used to infer the structure of proteins and RNA, predict the functional effects of clinical variants, design new proteins, etc. (Marks et al., 2011; Hopf et al., 2017; Frazer et al., 2020; Russ et al., 2020).

Standard methods that avoid MSA preprocessing. Although MSA preprocessing is problematic from the perspective of probabilistic modeling, the use of probabilistic models to infer multiple sequence alignments – that is, in order to *accomplish* the preprocessing – is standard. Perhaps the most widely used such method is the profile HMM, which, besides being used to infer multiple sequence alignments, is also at the core of modern sequence database search methods and is used to define sequence families, among many other applications (Durbin et al., 1998; Johnson et al., 2010; El-Gebali et al., 2019). In Section 4.2 we show that the MuE distribution generalizes a variety of popular methods including the profile HMM. While connections between various methods have been described before, the generalization offered by the MuE is both unified and comprehensive, delimiting the extent of the model class (Holmes, 2017). Note also that some of these models can be trained by interpreting an MSA as a point estimate of the latent alignment variable; this is distinct from the more common usage of MSA preprocessing described in Section 4.1 and is not subject to the same pathologies. The most closely related method to MuE observation models is the hidden Potts model (Wilburn & Eddy, 2020); we go further by providing a generalized approach to building and inferring similar models.

Natural language processing methods There has been intense recent interest in applying advances from natural language processing to biological sequences (Rives et al., 2019; Shin et al., 2021; Alley et al., 2019). The MuE is a type

Figure 3. The multiple sequence alignment of the initial dataset Y_1, Y_2 and Y_3 can change as more data, Y_4, is added.

of latent alignment model, a key model class in natural language processing; Deng et al. (2018) detail the close relationship between latent alignment and popular attention network methods. MuE observation models differ from standard latent alignment models in that (1) rather than regress on an observed sequence (e.g. a sentence in a language to be translated), the model regresses on a latent sequence X_i, and (2) the MuE is structured such that its latent alignment variable is interpretable as a *biological* alignment, not an alignment in the more generic sense used in natural language processing (Sections 4.2 and 4.3). Note that while the MuE itself is a relatively simple latent alignment model (an HMM), complex neural networks can be used to generate the latent sequence X_i; from a deep learning perspective, the MuE can be thought of as a biologically interpretable final layer.

4. Theory

4.1. Pathologies in MSA Preprocessing

MSA preprocessing is typically applied to static sequence datasets and used for parameter inference problems; its statistical pathologies emerge when when we attempt to predict unobserved or future sequences. To explain these pathologies, we focus on the i.i.d. case.[2] Consider the following modeling assumption, which is nearly universal in statistics:

Assumption 4.1 (I.i.d. data and model) *Let $p_0(x)$ be a probability distribution defined over a space $\mathcal{X}$, i.e. $p_0(x) \in \mathcal{P}(\mathcal{X})$ where $\mathcal{P}(\mathcal{X})$ is the set of all probability distributions over $\mathcal{X}$. We (1) assume that we observe independently and identically distributed samples $X_1, X_2, \ldots \sim p_0(x)$. In order to describe this process, we introduce a model $\mathcal{M} = \{q(x|\theta) : \theta \in \Theta\}$. We (2) assume $q(x|\theta) \in \mathcal{P}(\mathcal{X})$ for all $\theta \in \Theta$.*

Now consider models that use MSA preprocessing and take the following form, of which Equation 1 is a special case:

$$\text{Preprocess: } \{Y_{\text{MSA},1}, \ldots, Y_{\text{MSA},N}\} := f_{\text{MSA}}(\{Y_1, \ldots, Y_N\}),$$

$$\text{Model: } Y_{\text{MSA},i} \overset{iid}{\sim} p(y_{\text{MSA}}),$$

[2]Note that phylogenetic models, although not usually represented as i.i.d., are typically exchangeable and so possess an i.i.d. representation by de Finetti's theorem (Weinstein et al., 2020).

where $p(y_{\text{MSA}}) \in \mathcal{P}((\mathcal{B} \cup \{-\})^J)$. If we attempt to employ Assumption 4.1 to describe the preprocessed data $Y_{\text{MSA},1}, \ldots, Y_{\text{MSA},N}$ we see that it is violated. Part 1 of Assumption 4.1 fails because the preprocessed data cannot consist of independent observations: if a datapoint Y_{N+1} is added to the dataset, then past data, i.e. $Y_{\text{MSA},1}, \ldots, Y_{\text{MSA},N}$, can be altered (Figure 3). For instance, the new sequence may provide additional evidence to the MSA algorithm that sites in previously observed sequences are related to one another. Part 2 of Assumption 4.1 fails because the model is not defined over a space that encompasses future data: if a datapoint Y_{N+1} is added to the dataset, the value of J may change (Figure 3). For instance, the new sequence might be longer than any seen before. These failures occur on real sequence datasets, for typical values of N (Figure S2). Practically, the fact that MSA models violate Assumption 4.1 makes rigorous likelihood-based evaluation of their generalization capacity untrustworthy. If we do not know what space future data lives in, or how past data will be altered with future measurements, it is hard to trust that the average log likelihood of our model on a held out test set is genuinely reflective of future model performance. More technically, the violation of Assumption 4.1 causes standard justifications for the use of Bayes factors, heldout likelihood, prequential evaluation, etc. to fail, see e.g. Dawid (1984); Vapnik (1999); Dawid (2011).

Using MSA preprocessing also fails to account for uncertainty in the alignment (Wu et al., 2012; Toth-Petroczy et al., 2016). The goal of an MSA algorithm is to infer related sites among a set of sequences, but the resulting MSA is only a point estimate of this quantity.

4.2. Inferring Alignments

In this section we connect the MuE distribution to previously proposed probabilistic and non-probabilistic methods for inferring biological sequence alignments including MSAs, and describe how MuE observation models can be used to infer related sites and MSAs themselves. We start by more formally describing a biological pairwise alignment between two sequences X and Y, and then establish a connection with the latent state variable W in the MuE. Pairwise alignments serve as a diagrammatic representation of how two sequences X and Y may be related via insertion, deletion and substitution mutations.

Definition 4.2 (Biological pairwise alignment) *Let X and Y be sequences of length M and L respectively. A pairwise alignment $\mathcal{A}$ of X and Y with J columns is a matrix $[\mathcal{A}^{(x)}, \mathcal{A}^{(y)}]^\top$, where $\mathcal{A}^{(x)} \in (\mathcal{B} \cup \{-\})^J$ is a column vector of length J consisting of the letters of X, in order, and interspersed with gap symbols; similarly for $\mathcal{A}^{(y)}$. The alignment $\mathcal{A}$ must satisfy the condition that for every $j \in \{1, \ldots, J\}$ either $\mathcal{A}^{(x)}_j \in \mathcal{B}$ or $\mathcal{A}^{(y)}_j \in \mathcal{B}$ or both.*

Let j_l be the column of the alignment $\mathcal{A}$ in which the lth letter of Y falls, i.e. $\mathcal{A}^{(y)}_{j_l} = Y_l$ for $l \in \{1, \ldots, L\}$. Let g_l indicate whether the column j_l in $\mathcal{A}$ contains a gap, i.e. $g_l := \mathbb{I}(\mathcal{A}^{(x)}_{j_l} = -)$, where $\mathbb{I}(\cdot)$ is the indicator function which takes value 1 when the expression is true and 0 otherwise. Given X and Y, the sets $\{j_1, \ldots, j_L\}$ and $\{g_1, \ldots, g_L\}$ together uniquely define an alignment $\mathcal{A}$ (Remark S2.1). We can define a map from the latent state path W to a pairwise alignment $\mathcal{A}$ of X and Y.

Definition 4.3 (From latent states to biological alignments) *Given $W \sim p_{\mathrm{MuE}}(w|X, Y)$, let $g_l = \mathbb{I}(W_l > M)$ and $j_l = W_l - M g_l + \sum_{l'=1}^{l-1} g_{l'}$, for $l \in \{1, \ldots, L\}$. Note that this map is invertible.*

Under this definition, when $g_l = 0$, the letter Y_l is generated based on a letter X_{W_l} in the MuE, and Y_l and X_m are placed in the same column of the pairwise alignment $\mathcal{A}$; when $g_l = 1$, however, Y_l does not depend on X at all (it depends on c instead) and $\mathcal{A}^{(x)}_{j_l}$ has the gap symbol (Figure 2C).

A zoo of probabilistic and non-probabilistic methods have been proposed for inferring biological sequence alignments from data. Here we show that many of the most widely used methods can be unified as special case examples of the MuE which use Definition 4.3 to convert from W to $\mathcal{A}$.[3]

Proposition 4.4 (Unified) *For different choices of parameters c, ℓ, $a^{(0)}$, and $a^{(t)}$, (1) the Thorne-Kishino-Felsenstein model (Thorne et al., 1991), (2) the profile HMM, and (3) the conditional distribution of a sequence Y given a sequence X under the pair HMM (Durbin et al., 1998) are all special cases of the distribution $\mathrm{MuE}(X, c, \ell, a^{(0)}, a^{(t)})$, with a state-specific probability of the Markov chain terminating at each step. For another choice of parameters, the maximum a posteriori estimator $\hat{w} := \mathrm{argmax}_w p_{\mathrm{MuE}}(Y|X, w)$ corresponds to the Needleman-Wunsch alignment.*

See Section S2.2 for a proof. In the context of the profile HMM, point estimates of the latent alignment variables $W_1, \ldots, W_N$ associated with each observed sequence $Y_1, \ldots, Y_N$ are used to construct a multiple sequence alignment of the dataset by effectively merging pairwise alignments; sites in each Y_i generated by the same position in X are considered related, and placed in the same column. The same logic and algorithm can be applied to MuE observation models to define an MSA based on $W_1, \ldots, W_N$ (Figure 2D; Section S2.3).

The MuE offers not only a unified but also a comprehensive framework in the sense that HMMs which fail to satisfy

[3] So far we have not specified a model for the length L of the sequence Y. In the following proposition, we assume that there is some probability of the latent Markov chain terminating after each step l, and that this probability depends on the current state W_l.

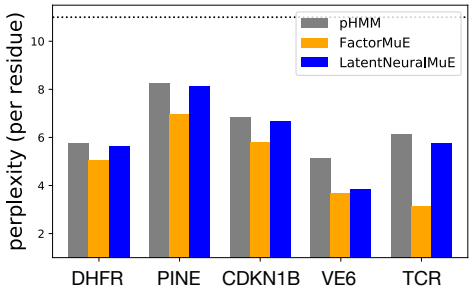

Figure 4. Predictive performance on a randomly heldout test set. Dotted line marks theoretically expected performance of the substitution matrix BLOSUM62 as a reference point (Section S5).

Constraint 2.2 cannot be interpreted, using Definition 4.3, as biological alignments (proof in Section S2.4):

Proposition 4.5 (Comprehensive) *Consider the setup of Definition 4.3 and assume each latent state $k \in \{1, \ldots, K\}$ of the MuE is Markov accessible under $a^{(0)}$ and $a^{(t)}$ (meaning that it can be reached with non-zero probability). Condition 2.2 is both necessary and sufficient to guarantee that with probability 1, W defines a valid pairwise alignment of X and Y via Definition 4.3.*

4.3. Comparison to Natural Language Models

Latent alignment models are used in natural language processing, often in combination with hard attention methods for inference (Deng et al., 2018). We can compare the MuE directly with a classic latent alignment model for statistical translation. The Vogel et al. (1996) model takes the form of a MuE model where X and Y are sentences in different languages, except that Condition 2.2 is violated (Section S2.5). As a result latent alignments are allowed to "double back" and rearrange the ordering of words in the regressor sentence X to generate Y.

5. Experiments

5.1. Predictive Performance

We have seen that models that use MSA preprocessing cannot be rigorously evaluated for their ability to predict sequences. In this section we empirically compare the predictive performance of MuE observation models to a standard model that possesses the same latent alignment structure, the profile HMM (pHMM) (Proposition 4.4).

Survey We started by examining five datasets of related protein sequences, ranging in size from 1,000 to 10,000 sequences (Section S6.1). Four were taken from non-redundant sequence databases: sequences similar to dihydrofolate reductase (DHFR), serine recombinase (PINE), cyclin dependent kinase inhibitor 1B (CDKN1B) and the human papillomavirus E6 protein (VE6) (Hopf et al., 2017; Toth-Petroczy et al., 2016; Tamarozzi & Giuliatti, 2018).

Table 1. Heldout perplexity on patient immune repertoire samples (each with 6,000 to 20,000 sequences). MS: multiple sclerosis. HC: healthy control. HC 1 consists of B cell receptors, the rest T cell receptors.

Dataset	HC 1	HC 2	HC 3	MS 1	MS 2	MS 3
pHMM	4.29	3.59	3.56	3.59	3.47	3.54
ICAMuE	**2.87**	**2.33**	**2.34**	**2.45**	**2.19**	**2.26**

The fifth dataset consisted of human T cell receptor (TCR) sequences from a healthy donor, obtained using single cell sequencing.

We extended probabilistic PCA and VAE models using the MuE observation distribution; we refer to these models as "FactorMuE" and "LatentNeuralMuE" respectively (model architectures are detailed in Section S3). We used stochastic variational inference, estimating the ELBO gradient using automatic differentiation, the reparameterization trick, and an inference network, and optimizing with Adam (Kucukelbir et al., 2017; Kingma & Welling, 2014; Rezende et al., 2014; Kingma & Ba, 2015). We evaluated model performance on a randomly held out 10% of sequences, quantified in terms of per residue (that is, per letter) perplexity (Section S5). The results show that FactorMuE models offer a consistent improvement over the standard pHMM model in every dataset, with an average change in perplexity of -1.50 and log Bayes factor $> 10^3$ across all datasets (Figure 4; Section S6.1). Meanwhile, the more complex LatentNeuralMuE model also improves over the pHMM in each dataset and overall (average perplexity change -0.42), but underperforms relative to the simpler FactorMuE model.

Patient immune repertoires We next explored further the application of MuE observation models to patient immune repertoire sequencing data, including both B and T cell receptors, taken from patients with autoimmune disease (multiple sclerosis) and healthy controls (Section S6.2) (Ramien et al., 2019). Understanding immune receptor repertoires is of crucial biomedical importance, but MSAs are considered highly untrustworthy when applied to this kind of data (see e.g. Figure S2). We extended another continuous model, an independent component analysis (ICA) model, with a MuE observation distribution ("ICAMuE"; Section S3.4). On a heldout 20% of data we find substantial improvements in perplexity over the pHMM across all six datasets (Table 1).

Disordered proteins Roughly $\sim$50% of the human proteome contains regions classified as disordered, but common bioinformatic pipelines are often considered highly untrustworthy when applied to disordered proteins because of uncertain MSAs. We examined 56 datasets, each consisting of sequences evolutionarily related to a disordered region of a human protein, that had been discarded in an MSA-based sequence modeling study (Toth-Petroczy et al., 2016). The study had sought in part to determine whether epistatic correlation occurred between amino acids at aligned sites (columns of the MSA), but was stymied in these particular datasets by highly uncertain MSAs. In a pHMM, conditional on a latent alignment W_i, the probability of observing a particular amino acid at a particular position in Y_i is independent of all other positions. In MuE observation models such as the FactorMuE, LatentNeuralMuE and ICAMuE, however, p_θ induces correlation between positions in Y_i conditional on W_i (Riesselman et al., 2018). To infer whether there is indeed epistatic correlation in a dataset, therefore, we can perform model selection, comparing a MuE observation model and a pHMM. Note that our approximate Bayesian inference procedure (for both models) integrates over all possible latent alignments, and that the pHMM is nested inside the MuE observation models in the sense of nested model selection (Dawid, 2011). We found that on 19 datasets an ICAMuE outperformed a pHMM at predicting a heldout 20% of sequences, finding evidence of epistatic correlation despite high alignment uncertainty; among these 19 datasets, the median perplexity decrease was 1.3 (Table S2, Section S6.3).

5.2. Learning Complex Biology

We examined further what the FactorMuE model had learned from a dataset of TCR sequences. T cell receptors are made up of two separate amino acid chains, α and β, which each develop according to a complex process of genome rearrangement termed V(D)J recombination, in which different V, D and J segments in the genome are, with some randomness and additional mutations, joined together with a constant region to produce a complete sequence (Figure 5A). We cross-referenced the latent representations of each sequence recorded in the dataset against supervised annotations of its segment types (Section S7). We found that the latent space is divided evenly in two, with one side containing TCRα sequences and one side TCRβ sequences (Figure 5B left). Each side contains clusters, which correspond with the type of V segment found in each TCR sequence (Figure 5B middle). The shorter J segments are found uniformly distributed across their corresponding α or β half, reflecting their ability to recombine with different V segments (Figure 5B right). See Section S7 for further results.

We next examined features learned by the FactorMuE model. In MuE observation models, we can separate out variation at conserved positions from variation produced by insertions and deletions by holding the latent alignment variable W_i fixed. In particular, we calculated

$$\nu_l := \left[\sum_{b=1}^{B} \left(\mathbb{E}[Y_{l,b} | \hat{w}_{\text{ref}}, z_1] - \mathbb{E}[Y_{l,b} | \hat{w}_{\text{ref}}, z_0] \right)^2 \right]^{1/2} \quad (3)$$

where the expectation is with respect to the variational ap-

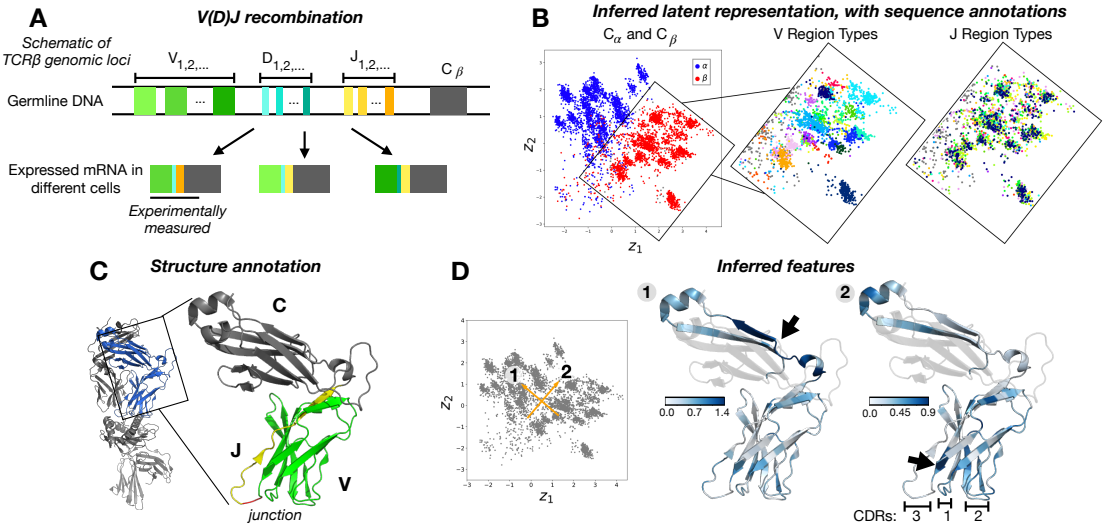

Figure 5. (A) Illustration of the TCRβ genomic locus; the TCRα locus is analogous, with C_α in place of C_β and no D segments (based on Abbas et al. (2018), Figure 8.7). (B) Inferred latent space representation of the TCR dataset, colored according to supervised annotations. Left: C_α and C_β chains. Middle: V types, $V_2, \ldots V_{30}$ (detailed legend in Figure S7). Right: J subtypes, $J_{1\text{-}1,\ldots,2\text{-}7}$ (detailed legend in Figure S7). (C) V (green), J (yellow) and constant C (gray) regions of the TCRβ chain in the reference structure PDB:2BNR, as well as V-J junction nucleotides (red) (Figure S7). (D) Projections ν of latent space vectors (left, in orange) into sequence space. Transparent areas correspond to the portion of the sequence that is not measured in the experiment. Arrows indicate peaks in ν.

proximation to the posterior, z_0 and z_1 are the head and tail of a vector in the latent space, $\hat{w}_{\text{ref}}$ is the maximum *a posteriori* estimate of W_{ref} based on a reference sequence Y_{ref}, and $l \in \{1, \ldots, L_{\text{ref}}\}$ where L_{ref} is the length of Y_{ref}. We plotted the vector ν on a TCR crystal structure for the reference sequence, and compared to a supervised annotation of the constant, V, D and J segments of the reference sequence (Figure 5CD). Consistent with the annotation of the latent representation, the vector normal to the hyperplane separating TCRα from TCRβ chains in the latent space (vector 1 in Figure 5D) primarily alters the sequence of the constant region, while the orthogonal vector (vector 2 in Figure 5D) primarily determines the sequence of the V segment. Along vector 2, the region of largest variation (the largest peak in ν_l) was the buried C-terminal end of the V segment, corresponding to the start of the CDR3 region, the key specificity-determining region of the receptor. Interestingly, even along vector 1 we observe high values of ν_l in the V segment, suggesting that there are systematic and heterogeneous differences between the V segment sequence distribution used in TCRα chains and in TCRβ chains (see Section S7 for further analysis).

5.3. Evolutionary Forecasting

We explored a novel application of generative probabilistic sequence models, evolutionary forecasting, which takes advantage of the capacity of MuE observation models to predict future sequences. Influenza A is responsible for an estimated 500,000 deaths a year and is an ongoing pandemic threat (Iuliano et al., 2018). It is also a model organism for understanding the dynamics of rapidly evolving pathogens, and forecasting its evolution is crucial in preparing vaccines and designing therapeutics (Luksza & Lässig, 2014; Laursen & Wilson, 2013). Previous forecasting methods have focused on predicting the relative fitness of existing strains in future years (Luksza & Lässig, 2014; Bush et al., 1999), or the antigenic properties of newly emerged strains (Neher et al., 2016; Hie et al., 2021). We instead predict the full amino acid sequence of the HA1 protein, the primary site of interaction with the immune system (Wiley et al., 1981). From the GISAID database we constructed a training set of influenza A(H3N2) HA1 sequences collected from patient samples from 1968 through 2013, and evaluated our predictions on sequences collected from 2014 through October 2019 (420 out of 2,042 sequences held out, 21% of the dataset) (Section S8) (Shu & McCauley, 2017). Insertions and deletions are considered rare, though not absent, in patient samples, so this dataset also offers an opportunity to evaluate MuE observation models in a distinct regime from that considered previously in Section 5.1.

As a benchmark we again used the pHMM, which can capture the observation that there exist key highly variable sites in the HA1 protein, an underlying motivation behind previous prediction methods such as Bush et al. (1999). We then incorporated sequence collection time as a covariate in new MuE observation models, using a linear regression model ("RegressMuE") and a neural network ("NeuralMuE") with MuE observation distributions (Section S3). The pHMM achieves a per residue perplexity of 1.32 and the RegressMuE improves this to 1.24 (log Bayes factor $> 10^3$; Figure 6A). This per residue perplexity difference corresponds

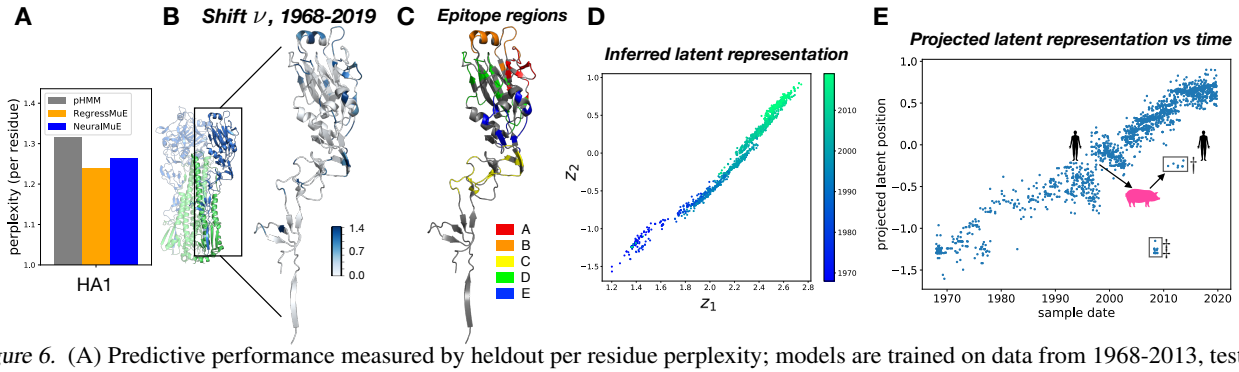

Figure 6. (A) Predictive performance measured by heldout per residue perplexity; models are trained on data from 1968-2013, tested on 2014-2020. (B) Magnitude of the shift in amino acid preference over time ν, for the RegressMuE, projected onto a reference HA1 structure (PDB:4O5N). The full hemagglutinin protein is shown on the left. (C) Classical epitope regions of the HA1 protein. (D) Inferred latent representation from a FactorMuE model, with sequences colored by the time at which the sample was collected (Section S8). (E) Y-axis: orthogonal projection of the latent representation of each sequence onto the least squares fit line relating z_1 and z_2. X-axis: time at which each sample was collected. Two clusters of outliers are marked by † and ‡.

to a factor of $\sim 10^{10}$ improvement in per sequence perplexity. The NeuralMuE has similar per residue perplexity (1.26) to the RegressMuE.

Next we investigated in detail what the model can tell us about how HA1 proteins have changed over time. We computed the magnitude of the shift in amino acid preference from 1968 to 2019 inferred by the model, with the latent MuE alignment variable kept fixed (quantified as ν_l, defined analogously to Equation 3 with times t_0 and t_1 replacing latent representations z_0 and z_1) (Figure 6B; Section S8). We found that sites with a large shift are often associated with antigenicity, consistent with the hypothesis that immune evasion is a key driver of influenza evolution. Residues that make up the classical epitope regions A-E of influenza show significantly larger shifts as compared to residues outside these regions (mean ν_l of 0.54 in epitopes A-E versus 0.09 in non-epitope sites, one sided Mann-Whitney U test $p < 10^{-18}$; Figures 6C and S12) (Wiley et al., 1981; Muñoz & Deem, 2005). The same observation holds for residues identified as key determinants of immune escape in recent high-throughput mutational antigenic profiling experiments (mean ν_l of 0.80 in sites with antigenic selection versus 0.24 elsewhere, one sided Mann-Whitney U test $p < 10^{-4}$; Section S8) (Lee et al., 2019).

The latent space representation of the influenza HA1 dataset learned by the FactorMuE model shows the data falling approximately along a line (Figure 6D; Section S8). The position of a sequence along this line is linearly proportional to the time at which the sequence was collected, though this information was not included in the model (correlation coefficient $\rho = 0.94$; Figure 6E) (Novembre & Stephens, 2008). Two clusters of outliers violate the proportionality rule. The first (marked by ‡) originated from mis-annotated entries in the GISAID database (Section S8). The second cluster (marked by †) appears in the early 2010s, but the latent representation of these sequences is close to that of sequences from the mid-1990s to early 2000s. Among this cluster of sequences, the ones that have been fully annotated were all collected from an outbreak in the United States of A(H3N2)v triple-reassortant viruses containing matrix protein genes from pandemic A(H1N1)pdm09. In 1998, A(H3N2)-derived viruses jumped from humans to swine, causing a large outbreak among swine, before recombining with other strains to produce this A(H3N2)v outbreak among humans in the 2010s (Jhung et al., 2013; Skowronski et al., 2012). The epidemiological history is consistent with our unsupervised latent representation, which shows that the cluster of outliers appearing in 2010-2013 most closely matches human samples last seen around 2000.

6. Discussion

MSAs are a powerful tool for analyzing biological sequences, but MSA preprocessing leads to statistical pathologies in generative models. MuE observation models offer a direct alternative to MSA preprocessing that does not abandon the underlying biological ideas that have made MSAs so successful. We hope that the MuE will enable rigorous application of a wide variety of new models and methodologies to biological sequence data.

Acknowledgments

We thank Chris Sander, John Ingraham, Smita Krishnaswamy, Alan Amin, Will Grathwohl and members of the Marks lab for discussion. We thank Elizabeth Wood for discussion and assistance with the T cell data, and Fritz Obermeyer and Eli Bingham for assistance with the Pyro implementation. We thank the anonymous reviewers for feedback and suggestions. ENW is supported by the Fannie and John Hertz Foundation. DSM is supported by the Chan Zuckerberg Initiative and an NIH TR01 grant (R01CA260415).

References

Abbas, A. K., Lichtman, A. H., and Pillai, S. *Cellular and Molecular Immunology*. Elsevier, ninth edition, 2018.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, 16(12):1315–1322, December 2019.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20(28):1–6, 2019.

Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., and Fitch, W. M. Predicting the evolution of human influenza A. *Science*, 286(5446):1921–1925, December 1999.

Dawid, A. P. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–292, 1984.

Dawid, A. P. Posterior model probabilities. In Bandyopadhyay, P. S. and Forster, M. R. (eds.), *Philosophy of Statistics*, volume 7, pp. 607–630. North-Holland, Amsterdam, January 2011.

Deng, Y., Kim, Y., Chiu, J., Guo, D., and Rush, A. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, pp. 9735–9747, 2018.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

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., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. The pfam protein families database in 2019. *Nucleic Acids Res.*, 47(D1):D427–D432, January 2019.

Felsenstein, J. *Inferring phylogenies*. Sinauer associates, Sunderland, MA, 2004.

Frazer, J., Notin, P., Dias, M., Gomez, A., Brock, K., Gal, Y., and Marks, D. Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning. December 2020.

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, December 2018.

Hie, B., Zhong, E. D., Berger, B., and Bryson, B. Learning the language of viral evolution and escape. *Science*, 371 (6526):284–288, 2021.

Holmes, I. H. Solving the master equation for indels. *BMC Bioinformatics*, 18(1):255, May 2017.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135, February 2017.

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., Cohen, C., Gran, J. M., Schanzer, D., Cowling, B. J., Wu, P., Kyncl, J., Ang, L. W., Park, M., Redlberger-Fritz, M., Yu, H., Espenhain, L., Krishnan, A., Emukule, G., van Asten, L., Pereira da Silva, S., Aungkulanon, S., Buchholz, U., Widdowson, M.-A., Bresee, J. S., and Global Seasonal Influenza-associated Mortality Collaborator Network. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet*, 391(10127):1285–1300, March 2018.

Jhung, M. A., Epperson, S., Biggerstaff, M., Allen, D., Balish, A., Barnes, N., Beaudoin, A., Berman, L., Bidol, S., Blanton, L., Blythe, D., Brammer, L., D'Mello, T., Danila, R., Davis, W., de Fijter, S., Diorio, M., Durand, L. O., Emery, S., Fowler, B., Garten, R., Grant, Y., Greenbaum, A., Gubareva, L., Havers, F., Haupt, T., House, J., Ibrahim, S., Jiang, V., Jain, S., Jernigan, D., Kazmierczak, J., Klimov, A., Lindstrom, S., Longenberger, A., Lucas, P., Lynfield, R., McMorrow, M., Moll, M., Morin, C., Ostroff, S., Page, S. L., Park, S. Y., Peters, S., Quinn, C., Reed, C., Richards, S., Scheftel, J., Simwale, O., Shu, B., Soyemi, K., Stauffer, J., Steffens, C., Su, S., Torso, L., Uyeki, T. M., Vetter, S., Villanueva, J., Wong, K. K., Shaw, M., Bresee, J. S., Cox, N., and Finelli, L. Outbreak of variant influenza A(H3N2) virus in the united states. *Clin. Infect. Dis.*, 57(12):1703–1712, December 2013.

Johnson, L. S., Eddy, S. R., and Portugaly, E. Hidden markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11:431, August 2010.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kingma, D. P. and Welling, M. Auto-Encoding variational bayes. In *International Conference on Learning Representations ICLR*, April 2014.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18(14):1–45, January 2017.

Laursen, N. S. and Wilson, I. A. Broadly neutralizing antibodies against influenza viruses. *Antiviral Res.*, 98(3): 476–483, June 2013.

Lee, J. M., Eguia, R., Zost, S. J., Choudhary, S., Wilson, P. C., Bedford, T., Stevens-Ayers, T., Boeckh, M., Hurt, A. C., Lakdawala, S. S., Hensley, S. E., and Bloom, J. D. Mapping person-to-person variation in viral mutations that escape polyclonal serum targeting influenza hemagglutinin. *Elife*, 8, August 2019.

Luksza, M. and Lässig, M. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, March 2014.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6(12):e28766, December 2011.

Muñoz, E. T. and Deem, M. W. Epitope analysis for influenza vaccine design. *Vaccine*, 23(9):1144–1148, January 2005.

Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A., and Shraiman, B. I. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl. Acad. Sci. U. S. A.*, 113(12):E1701–9, March 2016.

Novembre, J. and Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.*, 40(5):646–649, May 2008.

Ramien, C., Yusko, E. C., Engler, J. B., Gamradt, S., Patas, K., Schweingruber, N., Willing, A., Rosenkranz, S. C., Diemert, A., Harrison, A., Vignali, M., Sanders, C., Robins, H. S., Tolosa, E., Heesen, C., Arck, P. C., Scheffold, A., Chan, K., Emerson, R. O., Friese, M. A., and Gold, S. M. T cell repertoire dynamics during pregnancy in multiple sclerosis. *Cell Rep.*, 29(4):810–815.e4, October 2019.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, October 2018.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. April 2019.

Rush, A. M. Torch-Struct: Deep structured prediction library. February 2020.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369:440–445, 2020.

Särkkä, S. and García-Fernández, Á. F. Temporal parallelization of bayesian smoothers. *IEEE Trans. Automat. Contr.*, 66(1):299–306, 2020.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1): 2403, April 2021.

Shu, Y. and McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill.*, 22(13), March 2017.

Skowronski, D. M., Janjua, N. Z., De Serres, G., Purych, D., Gilca, V., Scheifele, D. W., Dionne, M., Sabaiduc, S., Gardy, J. L., Li, G., Bastien, N., Petric, M., Boivin, G., and Li, Y. Cross-reactive and vaccine-induced antibody to an emerging swine-origin variant of influenza a virus subtype H3N2 (H3N2v). *J. Infect. Dis.*, 206(12):1852–1861, December 2012.

Tamarozzi, E. R. and Giuliatti, S. Understanding the role of intrinsic disorder of viral proteins in the oncogenicity of different types of HPV. *Int. J. Mol. Sci.*, 19(1), January 2018.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22(22):4673–4680, November 1994.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.*, 25(24): 4876–4882, December 1997.

Thorne, J. L., Kishino, H., and Felsenstein, J. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33(2):114–124, August 1991.

Toth-Petroczy, A., Palmedo, P., Ingraham, J., Hopf, T. A., Berger, B., Sander, C., and Marks, D. S. Structured states of disordered proteins from genomic sequences. *Cell*, 167 (1):158–170.e12, September 2016.

Van Noorden, B. Y. R., Maher, B., and Nuzzo, R. Nature explores the most-cited research of all time. *Nature*, 514: 550–553, 2014.

Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.*, 10(5):988–999, 1999.

Vogel, S., Ney, H., and Tillmann, C. HMM-based word alignment in statistical translation. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, pp. 836–841, 1996.

Weinreb, C., Riesselman, A. J., Ingraham, J. B., Gross, T., Sander, C., and Marks, D. S. 3D RNA and functional interactions from evolutionary couplings. *Cell*, 165(4): 963–975, May 2016.

Weinstein, E. N., Frazer, J., and Marks, D. S. Deconvolving fitness and phylogeny in generative models of molecular evolution. In *Learning Meaningful Representations of Life Workshop at Neural Information Processing Systems*, 2020.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML)*, pp. 681–688, 2011.

Wilburn, G. W. and Eddy, S. R. Remote homology search with hidden potts models. *PLoS Comput. Biol.*, 16(11): e1008085, November 2020.

Wiley, D. C., Wilson, I. A., and Skehel, J. J. Structural identification of sites of Hong Kong influenza and their involvement in antigenic variation. *Nature*, 289, 1981.

Wu, M., Chatterji, S., and Eisen, J. A. Accounting for alignment uncertainty in phylogenomics. *PLoS One*, 7 (1):e30288, January 2012.