Non-identifiability and the Blessings of Misspecification in Models of Molecular Fitness and Phylogeny

Eli N. Weinstein[*1,2,5], **Alan N. Amin**[*2,3,5]
Jonathan Frazer[2], **and Debora S. Marks**[2,4,5]

[1] Program in Biophysics, Harvard University
[2] Department of Systems Biology, Harvard Medical School
[3] Program in Systems, Synthetic and Quantitative Biology, Harvard Medical School
[4] Broad Institute of Harvard and MIT
[5] Contact: eweinstein@g.harvard.edu, alanamin@g.harvard.edu, debbie@hms.harvard.edu
[*] These authors contributed equally

January 28, 2022

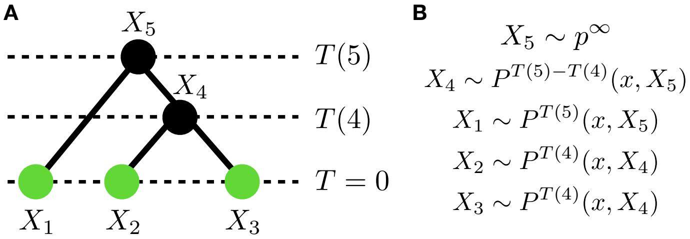
Abstract

Understanding the consequences of mutation for molecular fitness and function is a fundamental problem in biology. Recently, generative probabilistic models have emerged as a powerful tool for estimating fitness from evolutionary sequence data, with accuracy sufficient to predict both laboratory measurements of function and disease risk in humans, and to design novel functional proteins. Existing techniques rest on an assumed relationship between density estimation and fitness estimation, a relationship that we interrogate in this article. We prove that fitness is not identifiable from observational sequence data alone, placing fundamental limits on our ability to disentangle fitness landscapes from phylogenetic history. We show on real datasets that perfect density estimation in the limit of infinite data would, with high confidence, result in poor fitness estimation; current models perform accurate fitness estimation because of, not despite, misspecification. Our results challenge the conventional wisdom that bigger models trained on bigger datasets will inevitably lead to better fitness estimation, and suggest novel estimation strategies going forward.

1 Introduction

The past decades have witnessed a tremendous increase in the scale of genome sequence data available from across life. Recently, methods for estimating molecular fitness using generative sequence models have seen widespread success at translating this evolutionary data into predictions of the functional consequences of mutation. Such models have been shown to accurately predict the outcomes of experimental assays of protein function [Hopf et al., 2017, Riesselman et al., 2018, Meier et al., 2021], and have been applied to infer 3D structures of RNA and protein [Marks et al., 2011, Weinreb et al., 2016] and to design novel proteins [Shin et al., 2021, Russ et al., 2020, Madani et al., 2020]. The models have also been used to predict whether human mutations are pathogenic, directly informing the diagnosis of genetic disease [Frazer et al., 2021]. In this paper, we investigate how and why generative sequence models fit to evolutionary sequence data are successful at estimating molecular fitness, and how they might be improved and generalized going forward.

Existing approaches to fitness estimation with generative sequence models rest on an assumed relationship between density estimation and fitness estimation. Given a dataset of sequences $X_1, \ldots, X_N$, assumed to be drawn i.i.d. from some underlying distribution p_0, fitness models proceed by (1) fitting a probabilistic model q_θ to $X_{1:N}$ and (2) using the inferred density $\log q_{\hat{\theta}}(x) \approx \log p_0(x)$ as an estimate of the fitness $f(x)$ of a sequence x; this estimate in turn is used to predict other covariates such as whether the mutated sequence is pathogenic [Hopf et al., 2017, Riesselman et al., 2018, Frazer et al., 2021]. Innovation in fitness models has come out of a trend of building increasingly flexible models fit to increasing amounts of data: simple models



$$X_5 \sim p^{\infty}$$
$$X_4 \sim P^{T(5)-T(4)}(x, X_5)$$
$$X_1 \sim P^{T(5)}(x, X_5)$$
$$X_2 \sim P^{T(4)}(x, X_4)$$
$$X_3 \sim P^{T(4)}(x, X_4)$$

Figure 1: **Example JFPM for $N = 3$ observed sequences.** (A) An example phylogeny $\mathbf{H}$. (B) Generative process for sequences at each node of the phylogeny.

that treat each column of a sequence alignment independently were improved by energy-based models that accounted for epistasis [Hopf et al., 2017], which in turn were improved by deep variational autoencoders [Riesselman et al., 2018], which in turn were improved by deep autoregressive alignment-free models [Shin et al., 2021, Madani et al., 2020, Meier et al., 2021]. Naively, one might assume that these improvements have come from obtaining better and better estimates of the data distribution p_0, and improvements will continue with bigger models and bigger datasets. In this article, we argue that this presumption is incorrect.

Technical summary First, we show that that the true data distribution p_0 may not reflect fitness, and argue instead that we should be focused on estimating another distribution that does, p^{∞} (the "stationary distribution", to be defined below). In particular, we demonstrate that phylogenetic effects – i.e. the history of how current sequences evolved over time – can "distort" the observed data, leading to a situation where $p_0 \neq p^{\infty}$ (Sec. 2). Second, we show in this situation that p^{∞} and fitness f are non-identifiable: even with infinite data, there always exists some alternative fitness function $\tilde{f}$ that explains the same data just as well as f. This sets fundamental limits on what we can learn about fitness from evolutionary data (Sec. 3). Third, although exact estimation of p^{∞} is impossible, we show that it is still possible to get closer to p^{∞} than p_0, that is, to find a better estimator of fitness than the true data density p_0. This can be done by fitting to data a parametric generative sequence model $\mathcal{M} = \{q_\theta : \theta \in \Theta\}$ that is (approximately) well-specified with respect to p^{∞} (i.e. $p^{\infty} \in \mathcal{M}$) but *misspecified* with respect to the data distribution p_0 (i.e. $p_0 \notin \mathcal{M}$), thus illustrating the potential blessings of misspecification (Sec. 4). Fourth, we construct a hypothesis test to determine whether these blessings of misspecification occur on real data, with existing fitness estimation models; here, we rely on a Bayesian nonparametric sequence model to construct a credible set for p_0 (Sec. 6). Fifth, we apply our test to over 100 separate sequence datasets and fitness estimation tasks, to conclude that existing fitness estimation models systematically outperform the true data distribution p_0 at estimating fitness (Sec. 7). The takeaway is that better fitness estimation (i.e. better p^{∞} estimation) will not come from better density estimation (i.e. better p_0 estimation); bigger models and bigger datasets are not enough. Instead, better fitness estimation can come from developing models that describe p^{∞} better but the data density p_0 *worse*.

2 Models of Fitness and Phylogeny

In this section we show how p_0 may not accurately reflect the true fitness landscape, by developing a generative model of sequence evolution that takes into account both fitness and phylogeny. The model is general: it allows for arbitrarily complex epistatic fitness landscapes, and recovers standard generative phylogenetic and fitness models as special cases. Our concerns about the effects of phylogeny on fitness estimation are motivated by the widespread use – and trust – of phylogenetic models for evolutionary sequence data (phylogenetic models are far more widely applied than fitness models) [Hadfield et al., 2018, David and Alm, 2011, Felsenstein, 1985, 2004]. Although often inferred from the very same datasets, standard fitness models and standard phylogeny models make conflicting assumptions, which our general framework makes explicit.

Joint fitness and phylogeny models We define "joint fitness and phylogeny models (JFPMs)" using two elements: a description of how individual species (or populations or individuals) change over time, which depends on fitness f, and a description of the species' relationship to one another, a phylogeny $\mathbf{H}$. To describe the dynamics of individual species, let $P^{\tau}(x, x_0)$ denote the probability of sequence x_0 evolving into sequence x after time τ; in particular, $P^\tau(x, x_0)$ is assumed to be the transition probability of an irreducible continuous-time Markov chain defined over sequence space $\mathcal{X}$. For example, under neutral evolution (i.e. without selection based on fitness), $P^\tau(x, x_0)$ may follow a Jukes-Cantor model [Felsenstein, 2004]. With selection, for simple population genetics models (e.g. Moran or Wright processes), Sella and Hirsh [2005] demonstrate under general conditions that for any x_0,

$$P^\tau(x, x_0) \xrightarrow{\tau\to\infty} p^\infty = \frac{1}{\mathcal{Z}}\exp(\beta f(x)) \tag{1}$$

where $f(x)$ is the log fitness of the sequence x and $\beta > 0$ is a constant (Appx. A). The implication of Eqn. 1 is that the stationary distribution of the evolutionary dynamics follows a Boltzmann distribution, with energy proportional to the log fitness of the sequence. Estimating p^∞ is of interest because it provides a direct estimate of log fitness, up to a linear transform, since $f(x) = \beta^{-1}(\log p^\infty(x) + \log \mathcal{Z})$. (N.b. in the remainder of the paper, when we say "estimate fitness" we mean, implicitly, "estimate log fitness up to a linear transform".)

The sequences we observe, however, do not necessarily come from the stationary distribution. Instead, they are correlated with one another according to their evolutionary history. This is described by a phylogeny $\mathbf{H} = (V, E, T)$ consisting of a directed and rooted full binary tree with edges E and nodes V, along with time labels for the nodes, $T : V \to \mathbb{R}_+$ (Fig. 1A). Each node v is associated with a sequence X_v, drawn as $X_v \sim P^{\Delta t}(x, X_{v_0})$, where X_{v_0} is the sequence of the parent node, v is the child node, and $\Delta t = T(v_0) - T(v_1)$ is the length of the edge between them (Fig. 1B). The root sequence is drawn from p^∞. The observed datapoints $X_1, \ldots, X_N$ correspond to the leaf nodes. In general we will assume all leaves are observed at effectively the same time, the present day $T = 0$.

Simplifying assumptions Standard probabilistic phylogenetic models ignore fitness and assume

Assumption 2.1 (Pure phylogeny models (PMs)). *There is no difference in fitness among sequences, i.e. $f(x) = C$.*

Example models that fit this form include most of those used in BEAST [Drummond and Rambaut, 2007], MrBayes [Huelsenbeck and Ronquist, 2001], RaxML [Stamatakis, 2006], etc. Standard probabilistic fitness models, on the other hand, ignore phylogenetic history and assume that the stationary distribution has been reached,

Assumption 2.2 (Pure fitness models (FMs)). *Let τ_i be the distance in time between observed sequence X_i and its parent node. Take $\tau_i \to \infty$ for all i, which implies that*

$$X_i \overset{iid}{\sim} \frac{1}{\mathcal{Z}}\exp(\beta f(x)) \quad \text{for } i \in \{1, 2, \ldots\}. \tag{2}$$

The key implication of this assumption is that density estimation and fitness estimation are linked: the data follows $X_1, \ldots, X_N \sim_{iid} p_0 = p^\infty$, and so if we can estimate p_0 we can estimate the fitness. Example models include EVMutation [Hopf et al., 2017], DeepSequence [Riesselman et al., 2018], EVE [Frazer et al., 2021], etc. Note although Assumptions 2.1 and 2.2 do not conflict directly, conclusions made based on them conflict in practice: PMs typically infer finite and different lengths for branches (i.e. $\tau_i < \infty$), while FMs typically infer differences in fitness (i.e. $f(x) \neq C$), even when applied to the same dataset.

1D Example If Asm. 2.2 does *not* hold, then there is no reason for the distribution of observed sequences $X_1, X_2, \ldots$ to follow p^∞. We illustrate this with the most widely used example of a JFPM that does not use Assumptions 2.1 or 2.2: an Ornstein-Uhlenbeck tree (OUT) model [Felsenstein, 2004, Butler and King, 2004]. In this model, X is continuous, i.e. $X \in \mathbb{R}$, and evolves on a quadratic fitness landscape of the form $f(x) \propto (x - \mu)^2 + C$ according to the dynamics $P^\tau(x, x_0) = \text{Normal}\left(x_0 e^{-\frac{1}{2}\tau} + \mu, \sigma^2(1 - e^{-\tau})\right)$. The stationary distribution p^∞ is Normal(μ, σ^2). One can show (Appx. B.1) that for any phylogeny $\mathbf{H}$,

Proposition 2.3 (OUT observations). *The distribution of observed genotypes $X_{1:N}$ is drawn from a multivariate normal distribution with mean $\mu\vec{1}_N$ and covariance Σ where*

$$\Sigma_{ij} := \sigma^2 \exp(-\frac{1}{2}t_{ij}(\mathbf{H})) \quad \textit{for } i, j \in \{1, \ldots, N\}, \tag{3}$$

and $t_{ij}(\mathbf{H})$ is the total time of the shortest path between leaves i and j along the phylogeny $\mathbf{H}$.

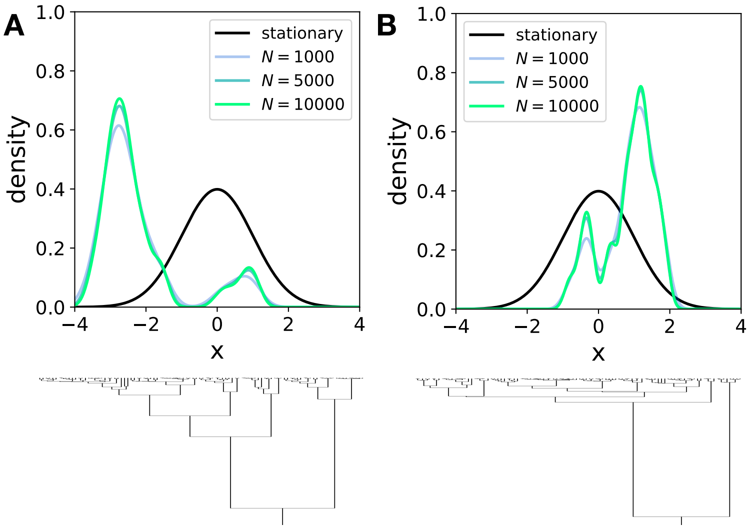


Figure 2: **Samples from an OUT.** (A) *Above:* Stationary distribution p^∞ and kernel density estimates of the distribution of samples p_0 from an OUT model for increasing N. *Below:* A subset of the phylogeny. (B) Same as (A) for an independent sample of $\mathbf{H}$.

We drew samples from the OUT with a Kingman coalescent prior on $\mathbf{H}$ (Bertoin [2010], Def. 2.1) and plotted their density (Fig. 2A). Even as $N \to \infty$, the distribution of samples does not follow p^∞. Moreover, rerunning the process with a new sample from the prior yields a very different distribution of samples (Fig. 2B).

3 Non-identifiability

In this section we investigate whether, given infinite sequence data, it is possible to infer fitness f without Asm. 2.2, and conversely, whether it is possible to infer phylogeny $\mathbf{H}$ without Asm. 2.1. That is, we are interested in whether fitness and phylogeny are identifiable in JFPMs. We conclude they are not: given infinite data generated with any f and $\mathbf{H}$, there exists some alternative $\tilde{f}$ and $\tilde{\mathbf{H}}$, where $\tilde{\mathbf{H}}$ satisfies Asm. 2.2, that explains the data equally well.

Naively, this result may be surprising: in FMs, each sequence is drawn independently, i.e. $X_i \perp\!\!\!\perp X_j | \mathbf{H}, f$, while in JFPMs and PMs there is (in general) correlation between sequences, i.e. $X_i \not\perp\!\!\!\perp X_j | \mathbf{H}, f$. One might then hope that examining correlations between sequences would enable us to infer whether Asm. 2.2 holds. However, we can show that these correlations are uninformative due to a symmetry in phylogenetic models, exchangeability.

Assumption 3.1 (Exchangeability)**.** *Let $m(X_1, X_2, \ldots)$ denote the marginal probability of an infinite set of sequences $X_1, X_2, \ldots$ integrating over all phylogenies, i.e. $m(X_1, X_2, \ldots) = \int p(X_1, X_2, \ldots | \mathbf{H}) p(\mathbf{H}) d\mathbf{H}$. Then, for any permutation π of the integers,*

$$m(X_1, X_2, \ldots) = m(X_{\pi(1)}, X_{\pi(2)}, \ldots). \tag{4}$$

Exchangeability says that if we had observed the sequences in a different order, it would not change their probability. In general, models of sequences observed at the same time (i.e. the present day, $T = 0$) satisfy exchangeability; for instance, models with a Kingman coalescent prior are exchangeable [Bertoin, 2010, Drummond and Rambaut, 2007]. Exchangeability implies that fitness and phylogeny are not identifiable. In particular, even if $X_1, X_2, \ldots$ are generated from a JFPM with a finite branch length phylogeny $\mathbf{H}$, we can describe the same data just as well using a model with an infinite branch length phylogeny $\tilde{\mathbf{H}}$ (an FM):

Theorem 3.2 (Non-identifiability)**.** *Assume $X_1, X_2, \dots$ satisfy Assumption 3.1. Then with probability 1 there exists some function $\tilde{f}$ such that*

$$X_i \overset{iid}{\sim} p_0 = \frac{1}{\tilde{\mathcal{Z}}} \exp(\beta \log \tilde{f}(x)) \quad \textit{for } i \in \{1, 2, \dots\}.$$

Proof. Applying de Finetti's Theorem (Kallenberg [2002], Thm. 11.10), there almost surely exists a random measure G such that for $i \in \{1, 2, ...\}$, $X_i \overset{iid}{\sim} G$. Let $p_G(x)$ be the pmf of G (we assume x is a finite discrete sequence; we can also work with continuous genotypes assuming the pdf $p_G(x)$ exists). Set $\tilde{f}(x) = [p_G(x)]^{1/\beta}$. □

This result says that the observed sequences from an exchangeable JFPM, $X_1, X_2, \dots$, are precisely i.i.d. samples from some p_0. Although in the standard tree representation $X_i \not\perp\!\!\!\perp X_j | \mathbf{H}, f$, there must be some alternative description of the same process where $X_i \perp\!\!\!\perp X_j | \tilde{\mathbf{H}}, \tilde{f}$. Fitness and phylogeny are thus non-identifiable: data generated from a JFPM with fitness f and phylogeny $\mathbf{H}$ can be described just as well using $\tilde{f}$ and $\tilde{\mathbf{H}}$, and vice versa.

The biological intuition behind Thm. 3.2 is that if two sequences are similar to each other and distant from a third, they may be similar either because they are closely related (i.e. the distance τ to the most recent common ancestor is small) or because they are in a local maxima of the fitness landscape. Without further assumptions, we cannot tell the difference between these two explanations. The machine learning intuition is that evolution, as described by a JFPM, is in effect a Markov chain Monte Carlo process whose stationary distribution gives the fitness. However, the samples we observe may not be fully independent: each pair of samples was initialized from the same point (the most recent common ancestor), and the burn-in since that point may not be sufficiently long. Without independent samples, our estimate of the stationary distribution will be biased.

Fitness inference as hyperparameter inference While general, Thm. 3.2 is not constructive, and does not tell us what the distribution p_0 actually is, or how exactly it differs from p^∞. Thm. 3.2 leaves unclear how much we need to know to learn the fitness landscape: could we infer fitness f if we knew the parametric form of p^∞, i.e. if we had some model $\mathcal{M}$ and knew that $p^\infty \in \mathcal{M}$? What if we also knew the underlying phylogeny $\mathbf{H}$? In the long branch limit (Asm. 2.2), fitness is identifiable if $\mathbf{H}$ is known; if $\mathcal{M}$ is also known, learning fitness is a matter of inferring model parameters. In the limit where all the branch lengths in the phylogeny are zero, the distribution of observations from a JFPM reduces to $X_1 \sim p_\infty$ and $X_1 = X_2 = X_3 = \dots$. Here fitness is non-identifiable even if $\mathbf{H}$ and $\mathcal{M}$ are known; learning fitness is a matter of learning from a single sample. In the realistic intermediate branch length case, if $\mathbf{H}$ and $\mathcal{M}$ are known, we will show that learning fitness is essentially a matter of *hyperparameter* rather than *parameter* inference.

We demonstrate this last claim by approximating OUTs as Gaussian process latent variable models (GPLVMs), finding that fitness only appears as a hyperparameter of the GP. The GPLVMs have latent variables $Z_1, Z_2, \dots$ that lie on the hyperbolic plane $\mathbb{H}$, and use the Gaussian process kernel $k(\cdot, \cdot) = \exp(-d(\cdot, \cdot))$, where $d(\cdot, \cdot)$ is a distance metric over $\mathbb{H}$. Let $\mathcal{W}_1(\cdot, \cdot)$ be the Wasserstein metric for distributions over infinite matrices, i.e. over $\mathbb{R}^{\infty \times \infty}$, using the sup norm on matrices.

Theorem 3.3 (GPLVM approximation of OUT)**.** *Assume a prior over phylogenies $\mathbf{H}$ that is exchangeable in its leaves and where the minimum time between any pair of nodes is greater than $\eta > 0$ with probability 1. Define the leaf distance matrix $\nu_{ij} = \log(\frac{1}{2} t_{ij}(\mathbf{H}))$. For any $\epsilon > 0$, there exists a.s. a GPLVM of the form,*

$$\begin{aligned} &G \sim \mathcal{G}, \qquad s \sim \text{GaussianProcess}(\mu, \sigma^2 k(\cdot, \cdot)), \\ &Z_i \overset{iid}{\sim} G \quad \textit{for } i \in \{1, 2, \dots\}, \\ &X_i = s(Z_i), \end{aligned} \tag{5}$$

where G is a random measure over $\mathbb{H}$, such that $\mathcal{W}_1(p(\nu), p(\tilde{\nu})) < \epsilon$, where $\tilde{\nu}_{ij} = \log(d(Z_i, Z_j))$.

If $\mathcal{W}_1(p(\nu), p(\tilde{\nu})) = 0$, the OUT and GPLVM produce identical distributions over $X_1, X_2, \dots$ a.e..

The proof is in Appx. B.2, and uses the embedding of Sarkar [2012]. This result says that, by embedding phylogenies $\mathbf{H}$ in a metric space, we can approximate an OUT arbitrarily well with a GPLVM; as the Wasserstein bound gets smaller, the distribution of covariance matrices of the two models get closer. In the

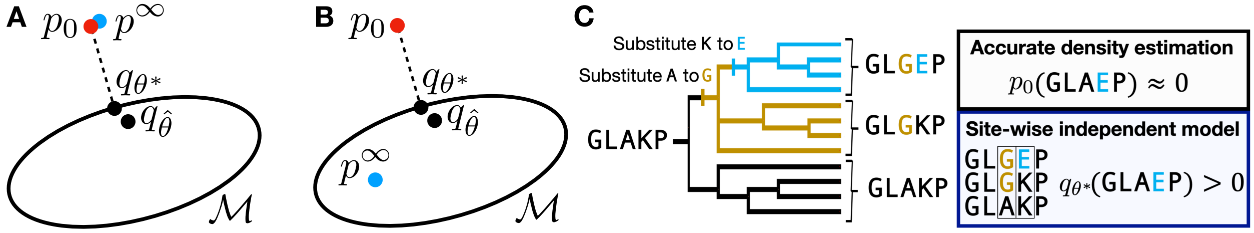


Figure 3: **Alternative explanations for the success of fitness estimation methods.** (A) Setup in which hypothesis 1 would hold true. (B) Setup in which hypothesis 2 would hold true. (C) Biological intuition for the blessings of misspecification (Hypothesis 2).

GPLVM, the observations are conditionally independent, $X_i \perp\!\!\!\perp X_j | s, G$, in line with Thm. 3.2. The phylogeny $\mathbf{H}$ enters the GPLVM only through the latent space embedding $Z_1, Z_2, \ldots$. Learning phylogeny, given the fitness landscape, is thus essentially a matter of inferring latent variables [Riesselman et al., 2018, Ding et al., 2019]. The fitness landscape enters the GPLVM only through the prior on the Gaussian process (i.e. through μ and σ). Inferring fitness given phylogeny is thus essentially a matter of inferring hyperparameters. This is both good and bad news for fitness inference. On the one hand, hyperparameters are often learned in practice, and doing so can yield substantially better predictions, so we should be able learn something about μ and σ given data (Williams and Rasmussen [2006], Chap. 5). On the other hand, hyperparameters are in general (though not always) non-identifiable, and therefore so is fitness [Mardia and Marshall, 1984]. Ho and Ané [2013] describe non-identifiability conditions for the OUT in particular. We conclude that even when $\mathbf{H}$ and $\mathcal{M}$ are known, fitness inference in JFPMs is fundamentally challenging.

4 Blessings of misspecification

We have demonstrated that phylogenetic effects can produce a data distribution p_0 that is not equal to the stationary distribution p^∞, and exact inference of p^∞ is in general impossible even with infinite data. Nonetheless, the practical success of fitness estimation methods suggest it is possible to at least approximate p^∞ from observational sequence data. Recall that existing methods proceed by fitting a probabilistic model $q_\theta \in \mathcal{M} = \{q_\theta : \theta \in \Theta\}$ to data $X_{1:N}$, typically via maximum likelihood estimation or approximate Bayesian inference, and then using the predicted log density $\log q_{\hat\theta}(x)$ as an estimate of the fitness of a sequence x. Why does this approach provide empirically successful estimates of p^∞? In this section we consider two hypotheses, either of which may hold true in theory. In Secs. 6-7 we develop and apply tests to evaluate them on real data.

Hypothesis #1 (informal). *Fitness estimation methods succeed by finding $q_{\hat\theta} \approx p_0$, since for all practical purposes on real data, $p_0 = p^\infty$.*

This hypothesis would make sense if Asm. 2.2 held, i.e. branch lengths were long enough in real datasets for $P^{\tau_i}(x, x_0)$ to be close to its stationary distribution. Under this explanation, better density estimators have been, and will continue to be, better fitness estimators. We should focus on developing models $\mathcal{M}$ that are well-specified with respect to the data, i.e. $p_0 \in \mathcal{M}$ (Fig. 3A).

Hypothesis #2 (informal). *Fitness estimation methods succeed by using models $\mathcal{M}$ that are misspecified with respect to p_0, i.e. $p_0 \notin \mathcal{M}$. The inferred model $q_{\hat\theta}$ is then closer to p^∞ than p_0 itself.*

To show this hypothesis is plausible, we prove that it is guaranteed to hold under general conditions. We study the projection of p_0 onto $\mathcal{M}$ via the Kullback-Leibler (KL) divergence, $q_{\theta^*} = \operatorname{argmin}_{q_\theta \in \mathcal{M}} \mathrm{KL}(p_0 \| q_\theta)$. The KL projection is relevant because maximum likelihood estimation minimizes the approximate KL divergence between the data and the model, and the posterior in Bayesian inference asymptotically concentrates around the maximum likelihood estimator [Miller, 2021]. We thus expect the fit model $q_{\hat\theta}$ to be close to q_{θ^*}, and get closer with N. Assume that $\mathcal{M}$ is "log-convex", meaning that for any $\theta, \theta' \in \Theta$ and $0 < r < 1$, there exists some θ'' such that $q_{\theta''}(x) = q_\theta(x)^r q_{\theta'}(x)^{1-r} / \sum_x q_\theta(x)^r q_{\theta'}(x)^{1-r}$; examples of log-convex models include the Potts model, as well as all other exponential family models.

Theorem 4.1 (Blessings of misspecification)**.** *Assume that the model $\mathcal{M}$ is log-convex and well-specified with respect to the stationary distribution, i.e. $p^\infty \in \mathcal{M}$. Assume q_{θ^*} exists and is unique. Then, if the model is misspecified with respect to the data distribution, i.e. $p_0 \notin \mathcal{M}$, we have*

$$\mathrm{KL}(q_{\theta^*} \| p^\infty) < \mathrm{KL}(p_0 \| q_{\theta^*}) + \mathrm{KL}(q_{\theta^*} \| p^\infty) \leq \mathrm{KL}(p_0 \| p^\infty). \tag{6}$$

But if the model is well-specified, i.e. $p_0 \in \mathcal{M}$, we have

$$\mathrm{KL}(q_{\theta^*} \| p^\infty) = \mathrm{KL}(p_0 \| p^\infty). \tag{7}$$

Proof. For part 1, apply Thm. 1 from Csiszar and Matus [2003]. For part 2, note that $q_{\theta^*} = p_0$ when $p_0 \in \mathcal{M}$. □

In words, the model projection $q_{\theta*}$ is closer to p^∞ than p_0 so long as as the model $\mathcal{M}$ is misspecified with respect to p_0 (Fig. 3B). To understand the biological intuition behind this result, consider a situation where two neutral mutations with no effect on fitness occur successively at different sites (Fig. 3C). Due to phylogenetic correlation, there is no observed sequence x^* in which the second mutation is present but not the first, so an accurate density estimator will find $p_0(x^*) \approx 0$. However, if we can guess correctly that the fitness landscape is independent across sites, then fitting a site-wise independent model $\mathcal{M}$ will imply the mutation is allowed, $q_{\theta^*}(x^*) > 0$, correctly inferring $p^\infty(x^*) > 0$.

Under Hypothesis 2, progress in the field of fitness estimation has *not* come from building better density estimators (Hypothesis 1), but rather from an iterative process of (1) hypothesizing, based partly on biophysical knowledge, models that are (approximately) well-specified with respect to p^∞ but not too flexible, such that $p_0 \notin \mathcal{M}$ and then (2) comparing their density estimates against experimental fitness measurements. We will show that on real data, Hypothesis 1 can often be rejected in favor of Hypothesis 2.

5 Related Work

Efforts to account for the effects of phylogeny in fitness estimation have a long history [Lapedes et al., 1999]. Practical generative sequence models that explicitly account for both epistatic fitness landscapes and phylogeny have long been sought, but stymied primarily by computational challenges [Ingraham, 2018, Rodriguez Horta et al., 2019]. In their place, a variety of non-generative (and often heuristic) methods for correcting for phylogeny have been proposed, including data reweighting schemes [Marks et al., 2011, Rodriguez Horta et al., 2019], data segmentation schemes [Colavin et al., 2022], post-inference parameter adjustments [Dunn et al., 2008], covariance matrix denoising methods [Qin and Colwell, 2018], simulation based statistical testing [Rivas et al., 2017], and more. In this article, we show that deconvolving fitness and phylogeny is not just computationally hard, but also in general statistically impossible: fitness and phylogeny are non-identifiable. We further show that use of a misspecified parametric model can on its own (without further corrections) partially adjust for phylogenetic effects.

Our results also intersect with the literature on robust statistics: we can think of the observed data distribution p_0 as a “distorted” version of the true distribution of interest p^∞. However, in typical robust inference frameworks (e.g. Huber’s epsilon contamination model), the observed distribution differs from the true distribution by the addition of outliers [Huber, 1992, Steinhardt, 2018]. In our setup, on the other hand, inliers are deleted, as phylogenetic correlations can mean the effective support of p_0 is *smaller* than that of p^∞ (Fig. 2).

6 Diagnostic Method

In this section, we develop diagnostic methods to discriminate between Hypothesis 1 and Hypothesis 2 (Sec. 4) based on observational sequence data and experimental fitness measurements, and validate these diagnostics in simulation. Recall that under Hypothesis 2, the estimate $q_{\hat{\theta}}$ from a parametric fitness model is a better estimate of fitness than the true data density p_0, while under Hypothesis 1, p_0 is better. Discriminating these two hypotheses on real data is nontrivial because we do not have access to p_0. Ideally, then, a diagnostic test would evaluate the probability that the true density p_0 outperforms $q_{\hat{\theta}}$ at predicting fitness, taking into

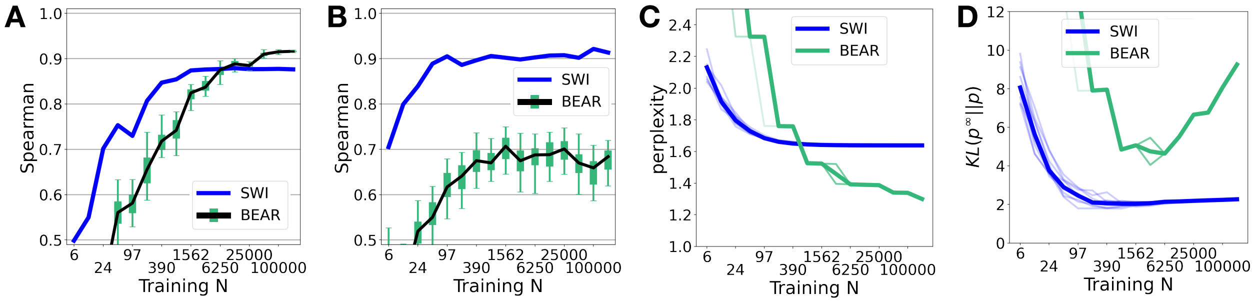


Figure 4: **The BEAR diagnostic applied to simulated data.** (A) Scenario 1. Spearman correlation between the maximum likelihood SWI model and the true fitness $\mathcal{S}_f(q_{\hat{\theta}})$, compared to the BEAR posterior distribution over $\mathcal{S}_f(p)$. Quantiles and 95% credible interval are shown with the green box and whisker plot. Points above (below) the whiskers correspond to SWI models that significantly outperform (underperform) the true data distribution. (B) Same as A, for Scenario 2. (C) Perplexity on heldout data of the BEAR and the SWI models in Scenario 2. Thick line corresponds to the average over 10 individual simulations (thin lines). (D) Same as C, comparing the KL divergence to p^{∞}.

account uncertainty in what p_0 could actually be, given the data. To accomplish this, we compute a posterior over p_0 using a Bayesian nonparametric sequence model. In particular, we apply the Bayesian embedded autoregressive (BEAR) model, which can be scaled to terabytes of data and satisfies posterior consistency (Amin et al. [2021], Thm. 35):

Theorem 6.1 (Summary of BEAR posterior consistency)**.** *Assume p_0 is subexponential, i.e. for some $t > 0$, $\mathbb{E}_{X\sim p_0}[\exp(t|X|)] < \infty$, where $|X|$ is the length of sequence X. Assume the conditions on the prior detailed in Amin et al. [2021]. If $X_1, X_2, \ldots \sim p_0$ i.i.d, then for $M > 0$ sufficiently large and $\epsilon \in (0, 1/2)$ sufficiently small,*

$$\Pi_{\text{BEAR}}(B(p_0, MN^{-\epsilon})|X_{1:N}) \xrightarrow{N\to\infty} 1$$

in probability, where $B(p, r)$ is a Hellinger ball of radius r centered at p, and $\Pi_{\text{BEAR}}(\cdot|X_{1:N})$ is the BEAR posterior.

Crucially, this result implies that the BEAR posterior will converge to effectively any value of p_0, no matter what p_0 is (unlike a parametric model's posterior). Moreover, BEAR quantifies uncertainty in its estimates, giving the range of possible values of p_0 that are consistent with the evidence.

We construct our diagnostic test by comparing the fitness estimation performance of $q_{\hat{\theta}}$ to the range of possible performances of p_0 estimated by BEAR. Let $\mathcal{S}_f(p)$ be a scalar score evaluating how accurately a density p predicts fitness f. In practice, $\mathcal{S}_f$ will be based on experimental and clinical measurements of quantities directly related to fitness.

Diagnostic test (Test Hypothesis 1 vs. Hypothesis 2.) *Hypothesis 1 $\mathcal{H}_1 : \mathcal{S}_f(q_{\hat{\theta}}) < \mathcal{S}_f(p_0)$. Hypothesis 2 $\mathcal{H}_2 : \mathcal{S}_f(q_{\hat{\theta}}) > \mathcal{S}_f(p_0)$. Accept Hypothesis 2 at significance level $\alpha > 0$ if*

$$\Pi_{\text{BEAR}}(\mathcal{S}_f(q_{\hat{\theta}}) > \mathcal{S}_f(p)|X_{1:N}) > 1 - \alpha. \tag{8}$$

Accept Hypothesis 1 at significance level α if

$$\Pi_{\text{BEAR}}(\mathcal{S}_f(q_{\hat{\theta}}) < \mathcal{S}_f(p)|X_{1:N}) > 1 - \alpha. \tag{9}$$

So long as $\mathcal{S}_f(p)$ is a well-behaved function of p (in particular, so long as $\mathcal{S}_f$ is continuous in a neighborhood of p_0 with respect to the topology of convergence in total variation), Thm. 6.1 implies that this diagnostic test will be asymptotically consistent, in the sense that it converges to the correct hypothesis in probability.

Simulations We next evaluate the performance of our diagnostic test on simulated data. We considered two scenarios, the first in which Hypothesis 1 holds, and the second in which Hypothesis 2 holds. In both, we let $\mathcal{M}$ be a site-wise independent (SWI) model, in which each position of the sequence is drawn independently, i.e. $X_l \sim \text{Categorical}(v_l)$ for $l \in \{1, \ldots, |X|\}$. The parameter v_l is in the simplex Δ_B, where $B + 1$ is the

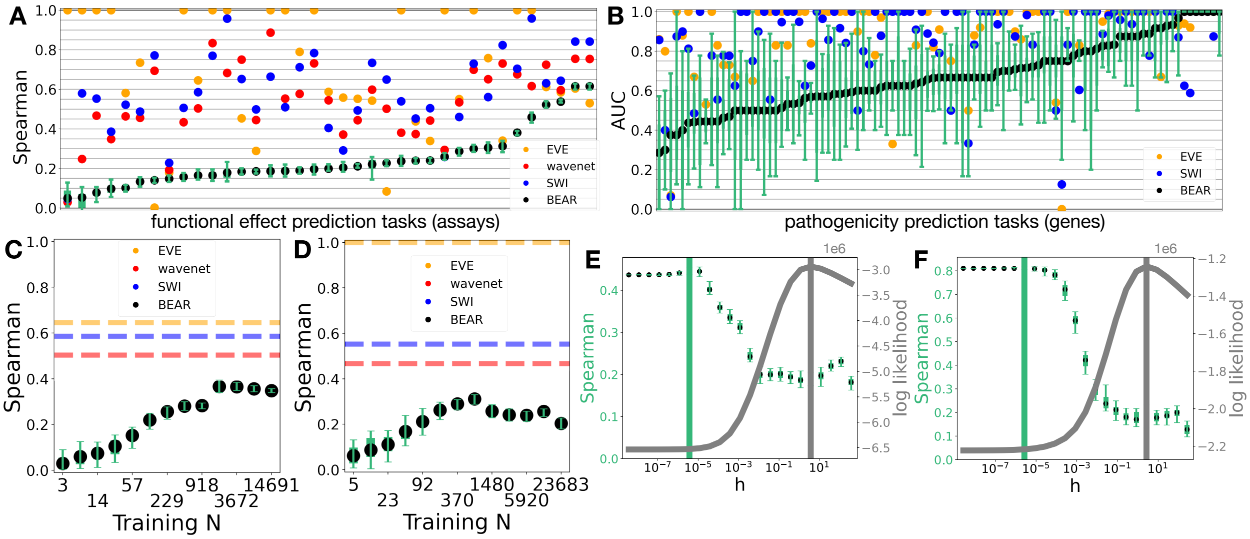


Figure 5: **Fitness estimation models systematically outperform the data distribution.** (A) Results for the first prediction task, predicting functional measurements in experimental assays. Quantiles and 95% credible interval of the BEAR posterior are shown with the green box and whisker plot. Points above (below) the whiskers correspond to fitness estimation models that significantly outperform (underperform) the true data distribution. (B) Results for the second prediction task, predicting variant pathogenicity in human genes. (C) Convergence of the BEAR posterior with datapoints N, for an example assay (β-lactamase). (D) Same as C, for another example assay (TIM barrel). (E) BEAR posterior Spearman (black and green) versus BEAR log likelihood (gray), interpolating between parametric and nonparametric regimes (low and high h), for an example assay (another β-lactamase assay). Peak Spearman indicated with vertical green line, peak log likelihood with gray. (F) Same as E, for another example assay (GAL4 DNA-binding domain).

alphabet size. (Further details in Appx. C.) In Scenario 1, the true data are generated according to a Potts model and $p_0 = p^\infty$. In this scenario, the SWI model is misspecified, and misspecification is *bad*: using a more flexible model will produce an asymptotically more accurate estimate of p^∞. We find that our diagnostic test asymptotically correctly accepts Hypothesis 1, in line with Thm. 6.1 (Figs. 4A and S3A). In Scenario 2, the true data are generated according to a JFPM with finite branch lengths, and $p^\infty \in \mathcal{M}$ while $p_0 \notin \mathcal{M}$. The mutational dynamics P^τ follow the Sella and Hirsh [2005] process. The phylogeny $\mathcal{H}$ is drawn from a Kingman coalescent. In this scenario, the SWI model is again misspecified, but misspecification is *good*: while the nonparametric BEAR model can achieve better density estimates than the SWI model (Fig. 4C), the SWI model outperforms BEAR at fitness estimation (Figs. 4D and S4). We find that our diagnostic test correctly accepts Hypothesis 2 (Figs. 4B and S3B).

A possible point of concern is that the test is poorly calibrated from a frequentist perspective, and in the low N regime accepts Hypothesis 2 in Scenario 1 more than 100α% of the time when the data is resampled from p_0 (Fig. S5A). This behavior is common in nonparametric Bayesian tests, and not necessarily a problem: the test is still valid from a purely Bayesian perspective. Nevertheless, on real data we will check that we are close to the large N regime by (1) checking that the BEAR posterior predictive is at least as close to p_0 as $q_{\hat{\theta}}$ is (as measured by perplexity on held out data; Figs. 4C and S5B) and (2) examining the plot of the BEAR posterior over $\mathcal{S}_f(p)$ as a function of N (as in Fig. 4AB), to check that it has converged.

7 Empirical Results

We now evaluate whether existing fitness estimation methods outperform the true data density p_0, i.e. whether we can reject Hypothesis 1 in favor of Hypothesis 2 on real data.

Tasks We consider two key prediction tasks where fitness models are applied in practice. The first task is to predict whether variants of a protein are functional, according to an experimental assay of protein function;

the metric $\mathcal{S}_f(\cdot)$ is the Spearman correlation between $p(x)$ and the assay result [Hopf et al., 2017]. There are typically ∼1000s of measurements per assay. The second task is to predict whether a variant of a protein observed in humans causes disease, according to clinical annotations; the metric $\mathcal{S}_f(\cdot)$ is the area under the ROC curve when $p(x)$ is used to predict whether or not a variant is pathogenic [Frazer et al., 2021]. There are typically only a handful of labels for each gene. For the first task, we considered 37 different assays across 32 different protein families, and for the second task, 97 genes across 87 protein families; for each protein family, we assembled datasets of evolutionarily related sequences, following previous work. Note that across the 37 assays and 97 genes, the data used for $\mathcal{S}_f$ comes from different experiments and different clinical evidence, often collected by different laboratories or doctors. As a consequence, our overall conclusions should be robust to the choice of $\mathcal{S}_f$.

Models We considered three existing fitness estimation models: a site-wise independent model (SWI), a Bayesian variational autoencoder (EVE [Frazer et al., 2021], which is similar to DeepSequence [Riesselman et al., 2018]), and a deep autoregressive model (Wavenet) [Shin et al., 2021]. Note that SWI and EVE, unlike Wavenet, require aligned sequences as training data. Details in Appx. D.

Results Applied to the first prediction task, our diagnostic test accepts Hypothesis 2 at significance level $\alpha = 0.025$ in 35/37 assays (95%) for SWI, 33/37 assays (89%) for EVE, and 36/37 assays (97%) for Wavenet (Fig. 5A). Applied to the second prediction task, our diagnostic test accepts Hypothesis 2 at significance level $\alpha = 0.025$ in 31/97 genes (32%) for SWI and 46/97 genes (47%) for EVE (Fig. 5B). Thus, fitness estimation models are capable of outperforming the true data distribution p_0. We found evidence for Hypothesis 1 in only a handful of examples: on the first task, Hypothesis 1 was accepted at significance level $\alpha = 0.025$ in 0/37 assays for SWI, 3/37 assays (8%) for EVE, and 0/37 assays for Wavenet, while on the second task, Hypothesis 1 was accepted for 5/97 genes (5%) for SWI and 4/97 genes (4%) for EVE. We confirmed that the diagnostic test was in the large N regime: BEAR outperformed Wavenet at density estimation, providing better predictive performance on 27/37 assays (73%) and similar performance on the remaining 10 assays (Fig. S6).[1] Example plots of the BEAR posterior's convergence with N on the first prediction task showed convergence to values of $\mathcal{S}_f$ well below that for parametric fitness estimation models (Figs. 5C and S7-S8). Overall, we conclude that there is strong evidence that existing fitness estimation methods reliably outperform the true data distribution p_0 across a range of datasets and tasks.

To study the tradeoffs between density estimation and fitness estimation in more depth, we smoothly and nonparametrically relaxed a parametric autoregressive (AR) model (Appx. D.4). We embedded the AR model (a convolutional neural network) into a BEAR model, and fit the BEAR model with empirical Bayes. We found evidence that the AR model was misspecified on every dataset, following the methodology of Amin et al. [2021]: the optimal h selected by empirical Bayes was on the order of $1 - 10$ in each dataset. Now, in the limit as the hyperparameter $h \to 0$, the BEAR model collapses to its embedded AR model; so by scanning h from low to high values we can interpolate between the parametric and nonparametric regime. We find a smooth tradeoff between $\mathcal{S}_f(p)$ and the likelihood of the data under the BEAR model, with higher h corresponding to better density estimation but worse fitness estimation (Fig. 5EF and S9). This relationship held across many datasets: the diagnostic test, evaluated against the AR model (the $h \to 0$ limit), accepts Hypothesis 2 in 28/37 assays (76%), but Hypothesis 1 in only 6/37 (16%) (Fig. S10). These results confirm that making a model well-specified (relaxing from a parametric to a nonparametric model) can bring improved density estimation at the cost of worse fitness estimation.

8 Discussion

In this article, we have argued that better density estimation does not necessarily lead to better fitness estimation. Our results changes the outlook for the future of fitness estimation: the common narrative that progress is inevitable through ever bigger models trained on ever bigger datasets appears to be false. Instead, progress will likely demand more fundamental methodological advances.

One future direction is to improve the current strategy of fitting misspecified models. For instance, it may be worthwhile to explore models that are *less* flexible than existing models and *worse* at density estimation, since they can increase the gap between $\mathrm{KL}(q_{\theta^*}\|p^\infty)$ and $\mathrm{KL}(p_0\|p^\infty)$ (Thm. 4.1). Another option is to improve the geometry of the model: while exponential family models are guaranteed to be log-convex (and thus

[1] Note that we cannot do this comparison for SWI or EVE since they are alignment-based [Weinstein and Marks, 2021].

can satisfy Thm. 4.1), we have no such guarantee for variational autoencoders or other neural network methods. Finally, uncertainty quantification is crucial for applications such as those in clinical genetics, but challenging in misspecified models [Szpiro et al., 2010, Miller and Dunson, 2019, Huggins and Miller, 2020]. Another future direction is to construct scalable JFPM models and carefully handle non-identifiability. Recent progress on amortized variational inference for phylogenetic models is promising [Vikram et al., 2019]. Non-identifiability is more challenging, and may require new assumptions and/or new methods of sensitivity analysis to infer the full set of fitness landscapes consistent with the data.

Finally, although this article has focused on technological applications of fitness models in solving prediction problems, fitness models also have implications for our fundamental understanding of evolution. Pure phylogeny models and pure fitness models present very different pictures of the past history of life: in PMs, similarities and differences among genetic sequences are determined primarily by history and ancestry (Asm. 2.1), while in FMs they are primarily determined by functional constraints (Asm. 2.2). PMs and FMs also present very different implications for the future of life: in PMs, the diversity of sequences seen in nature will likely expand dramatically going forward, while in FMs, the landscape of functional sequences has already been well-explored. Our results emphasize that where and to what extent each model offers an accurate picture of reality remains an open question.

References

A. N. Amin, E. N. Weinstein, and D. S. Marks. A generative nonparametric Bayesian model for whole genomes. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

J. Bertoin. Exchangeable coalescents. Nachdiplom Lectures, 2010.

M. A. Butler and A. A. King. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. *Am. Nat.*, 164(6):683–695, Dec. 2004.

A. Colavin, E. Atolia, A.-F. Bitbol, and K. C. Huang. Extracting phylogenetic dimensions of coevolution reveals hidden functional signals. *Sci. Rep.*, 12(1):820, Jan. 2022.

I. Csiszar and F. Matus. Information projections revisited. *IEEE Trans. Inf. Theory*, 49(6):1474–1490, June 2003.

L. A. David and E. J. Alm. Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328):93–96, Jan. 2011.

X. Ding, Z. Zou, and C. L. Brooks III. Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.*, 10(1):5644, Dec. 2019.

A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7:214, Nov. 2007.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

S. D. Dunn, L. M. Wahl, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb. 2008.

J. Felsenstein. Phylogenies and the comparative method. *Am. Nat.*, 125(1):1–15, Jan. 1985.

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

J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

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

L. S. T. Ho and C. Ané. Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. *Ann. Stat.*, 41(2):957–981, Apr. 2013.

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

P. J. Huber. Robust estimation of a location parameter. In S. Kotz and N. L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 492–518. Springer New York, New York, NY, 1992.

J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17 (8):754–755, Aug. 2001.

J. H. Huggins and J. W. Miller. Robust inference and model criticism using bagged posteriors. 2020.

J. Ingraham and D. Marks. Variational inference for sparse and undirected models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1607–1616. PMLR, 2017.

J. B. Ingraham. *Probabilistic Models of Structure in Biological Sequences.* PhD thesis, Harvard Medical School, 2018.

O. Kallenberg. *Foundations of Modern Probability.* Springer Science & Business Media, 2 edition, 2002.

M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, and D. R. Maglott. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, 46(D1):D1062–D1067, 2018.

A. S. Lapedes, B. G. Giraud, L. Liu, and G. Stormo. Correlated mutations in models of protein sequences: phylogenetic and structural effects. *Statistics in Molecular Biology, IMS Lecture Notes - Monograph Series*, 33:236–256, 1999.

A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher. ProGen: Language modeling for protein generation. Mar. 2020.

K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, Apr. 1984.

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

J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

J. W. Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *J. Mach. Learn. Res.*, 22(168):1–53, 2021.

J. W. Miller and D. B. Dunson. Robust bayesian inference via coarsening. *J. Am. Stat. Assoc.*, 114(527): 1113–1125, 2019.

C. Qin and L. J. Colwell. Power law tails in phylogenetic systems. *Proc. Natl. Acad. Sci. U. S. A.*, 115(4): 690–695, Jan. 2018.

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

E. Rivas, J. Clements, and S. R. Eddy. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods*, 14(1):45–48, Jan. 2017.

E. Rodriguez Horta, P. Barrat-Charlaix, and M. Weigt. Toward inferring potts models for phylogenetically correlated sequence data. *Entropy*, 21(11):1090, Nov. 2019.

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

R. Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane, 2012.

G. Sella and A. E. Hirsh. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U. S. A.*, 102(27):9541–9546, July 2005.

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

A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, Nov. 2006.

J. Steinhardt. *Robust Learning: Information Theory and Algorithms*. PhD thesis, Stanford University, 2018.

A. A. Szpiro, K. M. Rice, and T. Lumley. Model-robust regression and a Bayesian “sandwich” estimator. *Ann. Appl. Stat.*, 4(4):2099–2113, 2010.

S. Vikram, M. D. Hoffman, and M. J. Johnson. The LORACs prior for VAEs: Letting the trees speak for the data. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pages 3292–3301. PMLR, 2019.

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

E. N. Weinstein and D. S. Marks. A structured observation distribution for generative biological sequence prediction and forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 11068–11079. PMLR, 2021.

C. K. I. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

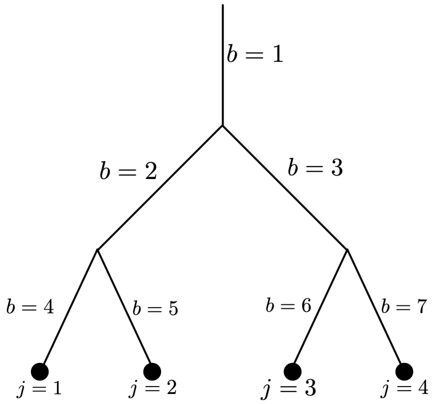


Figure S1: Tree labeling for the proof of Proposition 2.3

A Evolutionary dynamics models

Application of the Sella and Hirsh [2005] model (Eqn. 1) in JFPMs rests on a number of assumptions; we briefly the most relevant here.

When applying Eqn. 1 to amino acid sequences, as is typical for fitness estimation models, we ignore biases that come from the genetic code, which can modify the steady state probability of amino acids (in the absence of fitness effects) away from a uniform distribution. This is justified practically by the small effect sizes: if at steady state an amino acid has probability 1/64 instead of 1/20, the total difference in log probability is $\log(1/20) - \log(1/64) \approx 1$, which is small compared to (for instance) the log probability differences relevant for disease risk prediction with fitness models, which are ≈ 10 (Frazer et al. [2021], Extended data Fig. 3). Moreover, this bias only contributes an overall shift in amino acid probabilities, independent of position, and so does not change our main theoretical results. We ignore biases caused by asymmetric mutation rates for analogous reasons (though note they are often included in PMs in practice) [Sella and Hirsh, 2005].

The constant β depends on the effective population size, as well as the underlying population genetics model (Moran or Wright) and organismal ploidy (Sella and Hirsh [2005], Table 1). Following standard practice, we treat β as fixed for simplicity, though in reality it may vary over time and across lineages. Taking into account these possible changes clearly would not contradict our main theoretical result, that fitness and phylogeny are non-identifiable.

B Proofs

B.1 Proof of Proposition 2.3

N.b. this result is known in the literature (Ho and Ané [2013], Eqn. 1) but we are unaware of a proof, so we provide one here for completeness.

Proof. For notational convenience, we will work with a standardized OUT, with $\mu = 0$ and $\sigma = 1$. The final result can be obtained by translating and scaling the distribution of leaves. The transition distribution from point x' at time t' to point X at time t under the Ornstein-Uhlenbeck (OU) process is

$$X \sim \text{Normal}\left(x' e^{-\frac{1}{2}(t-t')}, 1 - e^{-(t-t')}\right). \tag{10}$$

This distribution can be reparameterized in location-scale form as

$$\begin{aligned}\epsilon &\sim \text{Normal}(0,1)\\ X &= x' e^{-\frac{1}{2}(t-t')} + \sqrt{1 - e^{-(t-t')}}\epsilon.\end{aligned}$$

As $t \to \infty$ we reach the stationary distribution Normal$(0,1)$. Let $b \in \{1, ..., \mathbf{B}\}$ index the branches of the tree, let λ_b be the length of branch b, and let $j \in \{1, ..., N\}$ index the leaves (observed species or sequences);

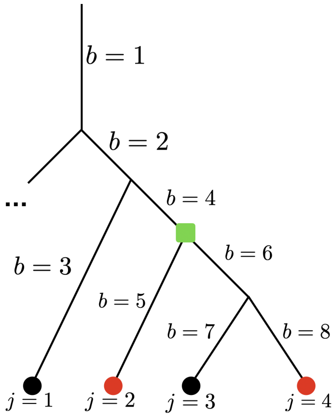


Figure S2: In red are the leaves considered in the examples in the proof of Proposition 2.3; in green is their most recent common ancestor.

see Fig. S1. We have assumed that the most recent common ancestor of the observed sequences was sampled from p^∞; this can be represented by adding a single branch length (indexed $b = 1$) to the root with length $\lambda_1 = \infty$. Let ϵ_b be the noise describing the OU diffusion over each branch. Let $\xi_{j,b}$ be the total time from leaf j to the nearest vertex on branch b, so long as branch b is on the path from leaf j to the root; otherwise, set $\xi_{j,b} = \infty$. For instance, in the diagram in Figure S1, we have $\xi_{1,4} = 0$, $\xi_{1,2} = \lambda_4$, $\xi_{1,1} = \lambda_4 + \lambda_2$, and $\xi_{1,5} = \xi_{1,6} = \xi_{1,7} = \xi_{1,3} = \infty$. We can now write the leaf position as

$$X_j = \sum_b e^{-\frac{1}{2}\xi_{j,b}} \sqrt{1 - e^{-\lambda_b}}\epsilon_b. \tag{11}$$

Define the matrix

$$M_{j,b} = e^{-\frac{1}{2}\xi_{j,b}} \sqrt{1 - e^{-\lambda_b}}, \tag{12}$$

such that $X_j = \sum_b M_{j,b}\epsilon_b$. We can now describe the complete leaf distribution as

$$\begin{aligned} \vec{\epsilon} &\sim \text{MultivariateNormal}(0, I_{\mathbf{B}}) \\ X_{1:N} &= M \cdot \vec{\epsilon}, \end{aligned}$$

where $I_{\mathbf{B}}$ is the $\mathbf{B}$-dimensional identity matrix. Thus, according to the location-scale representation of the multivariate normal,

$$X_{1:N} \sim \text{MultivariateNormal}(0, MM^\top). \tag{13}$$

We can simplify the covariance matrix $\Sigma := MM^\top$. First

$$\Sigma_{j,j'} = \sum_b M_{j,b}M_{j',b} = \sum_b e^{-\frac{1}{2}(\xi_{j,b}+\xi_{j',b})}(1 - e^{-\lambda_b}).$$

Before introducing the notation required to derive the general result, it's helpful to get a sense of how the derivation works; in the example tree (Figure S1),

$$\begin{aligned} \Sigma_{1,2} &= e^{-\frac{1}{2}(\lambda_4+\lambda_5)}(1 - e^{-\lambda_2}) + e^{-\frac{1}{2}(\lambda_4+\lambda_5+2\lambda_2)}(1 - e^{-\lambda_1}) \\ &= e^{-\frac{1}{2}(\lambda_4+\lambda_5)} + (-e^{-\frac{1}{2}(\lambda_4+\lambda_5+2\lambda_2)} + e^{-\frac{1}{2}(\lambda_4+\lambda_5+2\lambda_2)}) - e^{-\frac{1}{2}(\lambda_4+\lambda_5+2\lambda_2-2\lambda_1)} \\ &= e^{-\frac{1}{2}(\lambda_4+\lambda_5)}. \end{aligned}$$

The sum over b telescopes, leaving only the initial term, which corresponds to the total time between leaf node 1 and leaf node 2. To construct the general result, define $\tilde{b}_{j,j'}$ as the branch whose later node is the most recent common ancestor of leaves j and j'. In the example in Figure S2, $\tilde{b}_{2,4} = 4$. Let R be an ordered list of branches from $\tilde{b}_{j,j'}$ to $b = 1$, the earliest branch. In the example in Figure S2, $R = [4, 2, 1]$. Finally, let

$t_{jj'}$ be the length of the shortest path from leaf j to leaf j', the time from the most recent common ancestor to j plus the time to j'. In the example in Figure S2, $t_{2,4} = \lambda_5 + \lambda_6 + \lambda_8$. We now have

$$\begin{aligned}\Sigma_{jj'} &= \sum_{b=1}^{B} e^{-\frac{1}{2}(\xi_{j,b}+\xi_{j',b})}(1-e^{-\lambda_b}) \\ &= \sum_{b\in R} e^{-\frac{1}{2}(\xi_{j,b}+\xi_{j',b})}(1-e^{-\lambda_b}) \\ &= e^{-\frac{1}{2}t_{jj'}} - e^{-\frac{1}{2}(t_{jj'}+2\lambda_{\tilde{b}_{j,j'}})} + \sum_{k=2}^{|R|} e^{-\frac{1}{2}(t_{jj'}+2\sum_{k'=1}^{k-1}\lambda_{R_{k'}})}(1-e^{-\lambda_{R_k}}).\end{aligned}$$

Breaking down the telescoping sum, and using the fact that the final element of R is $t_1 = \infty$,

$$= e^{-\frac{1}{2}t_{jj'}} - e^{-\frac{1}{2}(t_{jj'}+2\sum_{k'=1}^{|R|}\lambda_{R_{k'}})} = e^{-\frac{1}{2}t_{jj'}}.$$

So we have the simple result that the covariance matrix depends just on the divergence times between leaves,

$$\Sigma_{jj'} = e^{-\frac{1}{2}t_{jj'}}. \tag{14}$$

Translating the distribution Eqn. 13 by μ and scaling by σ yields the result. □

B.2 Proof of Theorem 3.3

Before proving the result, we briefly clarify a definition in the statement of the theorem:

Definition B.1 (Exchangeable in leaves)**.** *Let* $\mathbf{H}$ *be a tree with countably infinite leaves and let* $\mathbf{H}_\pi$ *be a permutation of a phylogeny in its leaves, i.e. the same tree* $\mathbf{H}$ *with the leaves observed in a different order, according to a permutation* π*. A distribution over phylogenies is exchangeable in its leaves if* $p(\mathbf{H}) = p(\mathbf{H}_\pi)$ *for any permutation* π*.*

Proof. Outline: First, using the results from Sarkar [2012], we construct an embedding for each tree into the hyperbolic plane, being careful that the embedding preserves exchangeability. Second, we apply de Finetti's Theorem to obtain the conditionally independent representation of the joint distribution of $Z_1, Z_2, \ldots$*. Third, we use the distortion bound from Sarkar [2012] to bound the Wasserstein distance between* $p(\nu)$ *and* $p(\tilde{\nu})$*.*

First we describe the Sarkar [2012] $(1+\epsilon)$ distortion embedding algorithm setup. Vertices in phylogenetic trees have maximum degree three, and, by assumption, the minimum edge length in a tree $\mathbf{H}$ is greater than $\eta > 0$ with probability one. For any $\epsilon' > 0$, choose a $\rho < \pi/3$ and a scale factor

$$\lambda > \left(\frac{1+\epsilon'}{\epsilon'}\right)\frac{2k}{\eta}\log\tan\frac{\rho}{2}, \tag{15}$$

where k is the Gaussian curvature of the hyperbolic plane $\mathbb{H}$ (for most hyperbolic geometry models, and in particular the Lorentz manifold, $k = -1$). Then, let $h_1(\mathbf{H}), h_2(\mathbf{H}), \ldots$ be the position of the leaves in the embedding of $\mathbf{H}$ produced by the $(1+\epsilon)$ distortion embedding algorithm in Sarkar [2012], using edge scale factor λ, and ρ separated cones with cone angle $2\pi/3 - 2\rho$. Taking the last line of the proof of Theorem 6 in Sarkar [2012], we are guaranteed that even for a countably infinite number of leaves,

$$\begin{aligned}\max_{i,i'} \frac{\lambda t_{ii'}(\mathbf{H})}{\tilde{d}(h_i(\mathbf{H}), h_{i'}(\mathbf{H}))} &\leq 1+\epsilon' \\ \max_{i,i'} \frac{\tilde{d}(h_i(\mathbf{H}), h_{i'}(\mathbf{H}))}{\lambda t_{ii'}(\mathbf{H})} &= 1,\end{aligned} \tag{16}$$

where $i, i' \in \mathbb{N} := \{1, 2, \ldots\}$, and $\tilde{d}(\cdot,\cdot)$ is the hyperbolic distance function.

Next we will modify the embedding function h to ensure that the distribution of embedded leaves is exchangeable. Let $[\mathbf{H}]$ be the set of phylogenetic trees that are equivalent to $\mathbf{H}$ up to reordering of the vertices. For each equivalence class $[\mathbf{H}]$ we choose one ordering of the vertices to be the canonical tree $\hat{\mathbf{H}}([\mathbf{H}])$, and for any tree $\mathbf{H}$ let $\pi^c(\mathbf{H})$ be the leaf permutation such that the reordered tree $\mathbf{H}_{\pi^c(\mathbf{H})} = \hat{\mathbf{H}}([\mathbf{H}])$. Now define the modified leaf embedding function $h'(\mathbf{H}) := h_{\pi(\mathbf{H})}(\mathbf{H}_{\pi^c(\mathbf{H})})$ where $\pi(\mathbf{H})$ is the inverse permutation of $\pi^c(\mathbf{H})$. Since by assumption the prior $p(\mathbf{H})$ on the phylogenetic tree is exchangeable, we can rewrite $p(\mathbf{H})$ using the induced distribution over equivalence classes $p([\mathbf{H}])$ as

$$\begin{aligned}[\mathbf{H}] &\sim p([\mathbf{H}])\\ \pi &\sim \text{Permutation}\\ \mathbf{H} &:= \hat{\mathbf{H}}([\mathbf{H}])_\pi,\end{aligned}$$

where Permutation is the uniform distribution over all permutations of $\mathbb{N} := \{1, 2, \ldots\}$. We now define the distribution over leaf embeddings as

$$\begin{aligned}\mathbf{H} &\sim p(\mathbf{H})\\ Z_{1:\infty} &:= h'_{1:\infty}(\mathbf{H}),\end{aligned} \tag{17}$$

which we can rewrite as

$$\begin{aligned}[\mathbf{H}] &\sim p([\mathbf{H}])\\ \pi &\sim \text{Permutation}\\ Z_{1:\infty} &:= h_\pi(\hat{\mathbf{H}}([\mathbf{H}])).\end{aligned}$$

The distribution $p(Z_1, Z_2, ...)$ is therefore exchangeable. Applying de Finetti's Theorem [Kallenberg, 2002] we have a.s.

$$\begin{aligned}G &\sim \mathcal{G}\\ Z_i &\overset{iid}{\sim} G \text{ for } i \in \{1, 2, \ldots\}\end{aligned} \tag{18}$$

where G is a random measure distributed according to a prior $\mathcal{G}$. Moreover, the embedding distortion bounds (Eqn. 16) are preserved for each $\mathbf{H}$, since

$$\begin{aligned}1 + \epsilon \geq \max_{i,i'} \frac{\lambda t_{ii'}(\hat{\mathbf{H}}([\mathbf{H}]))}{\tilde{d}(h_i(\hat{\mathbf{H}}([\mathbf{H}])), h_{i'}(\hat{\mathbf{H}}([\mathbf{H}])))} &= \max_{i,i'} \frac{\lambda t_{\pi_i \pi_{i'}}(\mathbf{H}_{\pi^c(\mathbf{H})})}{\tilde{d}(h_{\pi_i}(\mathbf{H}_{\pi^c(\mathbf{H})}), h_{\pi_{i'}}(\mathbf{H}_{\pi^c(\mathbf{H})}))}\\ &= \max_{i,i'} \frac{\lambda t_{ii'}(\mathbf{H})}{\tilde{d}(h'_i(\mathbf{H}), h'_{i'}(\mathbf{H}))},\end{aligned} \tag{19}$$

and by the same logic

$$1 = \max_{i,i'} \frac{\tilde{d}(h_i(\hat{\mathbf{H}}([\mathbf{H}])), h_{i'}(\hat{\mathbf{H}}([\mathbf{H}])))}{\lambda t_{ii'}(\hat{\mathbf{H}}([\mathbf{H}]))} = \max_{i,i'} \frac{\tilde{d}(h'_i(\mathbf{H}), h'_{i'}(\mathbf{H}))}{\lambda t_{ii'}(\mathbf{H})}. \tag{20}$$

We will now construct the Wasserstein bound. Define the joint distribution over ν and $\tilde{\nu}$,

$$\begin{aligned}\mathbf{H} &\sim p(\mathbf{H})\\ \nu_{ii'}(\mathbf{H}) &:= \log(\frac{1}{2} t_{ii'}(\mathbf{H}))\\ \tilde{\nu}_{ii'}(\mathbf{H}) &:= \log(d(h'_i(\mathbf{H}), h'_{i'}(\mathbf{H})))\end{aligned} \tag{21}$$

where we have chosen $d(\cdot, \cdot) = \frac{1}{2\lambda}\tilde{d}(\cdot, \cdot)$. Note that the marginal distribution of ν matches its definition in the statement of the theorem, and that, applying Eqn. 17 and Eqn. 18, the marginal distribution of $\tilde{\nu}$ also matches its definition. Using the fact that log is a monotonically increasing function, Eqn. 19 gives

$$\begin{aligned}\log \sup_{i,i'} \frac{\exp(\nu_{ii'}(\mathbf{H}))}{\exp(\tilde{\nu}_{ii'}(\mathbf{H}))} &\leq \log(1 + \epsilon)\\ \sup_{i,i'}[\nu_{ii'}(\mathbf{H}) - \tilde{\nu}_{ii'}(\mathbf{H})] &\leq \epsilon,\end{aligned}$$

and similarly using the bound from Eqn. 20, $\sup_{i,i'}[\tilde{\nu}_{i,i'}(\mathbf{H}) - \nu_{i,i'}(\mathbf{H})] \leq 0$. Thus, with probability 1 under $p(\mathbf{H})$,

$$\|\nu(\mathbf{H}) - \tilde{\nu}(\mathbf{H})\|_\infty = \sup_{i,i'} |\nu_{ii'}(\mathbf{H}) - \tilde{\nu}_{ii'}(\mathbf{H})| \leq \epsilon.$$

Recall that the Wasserstein distance between the distribution of two random variables ν and $\tilde{\nu}$ can be written as

$$\mathcal{W}_1(p(\nu), p(\tilde{\nu})) = \inf_{\gamma \in \mathcal{J}} \mathbb{E}_\gamma[\|\nu - \tilde{\nu}\|_\infty]$$

where $\mathcal{J}$ is the set of joint distributions with marginals corresponding to the distributions of ν and $\tilde{\nu}$ (Dudley [2002], Chap. 11.8). Using the joint distribution in Eqn. 21, the Wasserstein distance is bounded by

$$\mathcal{W}_1(p(\nu), p(\tilde{\nu})) \leq \mathbb{E}_{\mathbf{H} \sim p(\mathbf{H})}[\|\nu(\mathbf{H}) - \tilde{\nu}(\mathbf{H})\|_\infty] \leq \epsilon. \tag{22}$$

Now consider the case where $\mathcal{W}_1(p(\nu), p(\tilde{\nu})) = 0$. (N.b. in this case, we do not need to assume that the minimum time between nodes in $\mathbf{H}$ is greater than $\eta > 0$.) Since the Wasserstein metric is a metric on the space of probability distributions (Dudley [2002] Lemma 11.8.3), $p(\nu) = p(\tilde{\nu})$ a.e.. Using the standard properties of Gaussian processes (Williams and Rasmussen [2006], Chap. 2), the GPLVM model (Eqn. 5) can be written as

$$\begin{aligned} G &\sim \mathcal{G} \\ Z_i &\overset{iid}{\sim} G \text{ for } i \in \mathbb{N} \\ \tilde{\nu}_{ii'} &:= \log d(Z_i, Z_{i'}) \\ X_{1:\infty} &\sim \text{MultivariateNormal}(\mu, \Sigma_{ii'} := \sigma^2 \exp(-\exp \tilde{\nu}_{ii'})), \end{aligned} \tag{23}$$

which is equivalent to the OUT distribution,

$$\begin{aligned} \mathbf{H} &\sim p(\mathbf{H}) \\ \nu_{ii'} &:= \log[\frac{1}{2} t_{ii'}(\mathbf{H})] \\ X'_{1:\infty} &\sim \text{MultivariateNormal}(\mu, \Sigma_{i,i'} := \sigma^2 \exp(-\exp \nu_{i,i'})). \end{aligned} \tag{24}$$

So the distribution $p(X_{1:\infty})$ produced by the GPLVM is equivalent to the distribution $p(X'_{1:\infty})$ produced by the OUT model a.e.. □

C Simulation Details

In both scenarios, we generated sequences of fixed length $|X| = 30$, with an alphabet size of $B + 1 = 4$ (corresponding to nucleotides).

Scenario 1 We simulated from a Potts model

$$p_{\text{POTTS}}(x) = \frac{1}{\mathcal{Z}} \exp\left(\sum_l \sum_b h_{lb} x_{lb} + \sum_l \sum_{l'>l} \sum_b \sum_{b'} e_{ll'bb'} x_{lb} x_{lb'}\right)$$

where h is the sitewise energies, e is the pairwise energies, x is a one-hot sequence encoding, l indexes sequence positions and b indexes letters. Following the simulations in Ingraham and Marks [2017], which were intended to roughly match the statistics of typical real protein Potts models, we drew $h_{lb} \sim \text{InvGamma}(2, 0.8)$ and

$$\begin{aligned} A_{ll'} &= \begin{cases} 1 & \text{if } l' = l + 1 \\ \text{Bernoulli}(0.1) & \text{otherwise} \end{cases} \\ B_{ll'bb'} &\sim \text{Normal}(0, 1.2) \\ e_{ll'bb'} &= A_{ll'} B_{ll'bb'}. \end{aligned}$$

The energies h and e were drawn once, and the same values used across independent simulations. We sampled from the model using a Gibbs sampler with 100 steps of burn-in and 10 parallel chains using the code from Ingraham and Marks [2017] (`https://github.com/debbiemarkslab/persistent-vi`). We shuffled the resulting samples to remove autocorrelation.

Scenario 2 We used a site-wise independent fitness function:

$$f(x) = \sum_{l=1}^{30} \sum_{b} h_{lb} x_{lb},$$

with site-wise residue biases h_l, where x_l is a one-hot encoding of the letter at the l-th position of x. To generate phylogenetically correlated sequences, we sampled phylogenetic trees from a Kingman Coalescent (Bertoin [2010], Def. 2.1) with rate 1. Starting from a random sequence drawn from the steady state distribution at the root, we evolved the sequence simulating a Wright process in a haploid population (Sella and Hirsh [2005], Eqn. 3) according to the tree and fitness function. In particular, for sequences x_0, x that are one mutation away, the mutation rate is

$$\lim_{\tau \to 0} \frac{1}{\tau} P^{\tau}(x, x_0) = N_{\text{eff}} \frac{e^{2(f(x)-f(x_0))} - 1}{e^{2N_{\text{eff}}(f(x)-f(x_0))} - 1},$$

where we set the effective population size to $N_{\text{eff}} = 10000$. This stochastic process has steady state

$$p^{\infty}(x) \propto \exp\left(2(N_{\text{eff}} - 1) f(x)\right),$$

(Sella and Hirsh [2005], Eqn. 7).

SWI model We fit the SWI model with maximum likelihood estimation.

BEAR model In these simulations, we used a vanilla BEAR model with a uniform embedded AR model (i.e. a Bayesian Markov model) for simplicity. We set the Dirichlet prior concentration to the constant $\alpha = 0.5$. Based on the theoretical analysis in Amin et al. [2021] (Thm. 35), we used a prior on lags of the form

$$p(L) \propto \exp(-B^L) \tag{25}$$

where B is the alphabet size (4 for nucleotides). We inferred the prior via empirical Bayes, marginalizing over the transition probabilities following the protocol in Amin et al. [2021]. Conditional on lag L, sampling from the posterior over the BEAR model is straightforward thanks to Dirichlet-Categorical conjugancy.

Evaluation We defined $\mathcal{S}_f$ following standard protocols for fitness estimation models. In particular, we let $\mathcal{S}_f(p)$ be the Spearman correlation between $p(x)$ and $f(x)$ for $x \in \Lambda$ where Λ consists of all possible single point mutations (i.e. single letter changes) of an initial ("wild-type") sequence. The wild-type sequence was chosen as the most likely sequence under p^{∞}, computed exactly for Scenario 2 and estimated based on the 10^6 samples for Scenario 1.

To estimate model perplexity (Fig. 4C and S5B), we used $N = 10,000$ independent sequences from p_0 and computed the per-residue perplexity

$$\exp\left(-\frac{1}{\sum_{n=1}^{N} |X_n|} \sum_{n=1}^{N} \log p(X_n)\right), \tag{26}$$

where $|X_n|$ is the sequence length and $p(X_n)$ is the probability of the sequence under the model.

To estimate the KL to the fitness distribution in Scenario 2 (Fig. 4D) , we sampled $N = 10,000$ independent sequences from p^{∞}, $\{X_1, \ldots, X_N\}$ and estimated

$$\text{KL}(p^{\infty}||p) \approx H(p^{\infty}) - \frac{1}{N} \sum_{n=1}^{N} \log p(X_n),$$

where $H(p^{\infty})$ is the entropy of p^{∞}, which can be computed analytically. For BEAR, we plotted the KL to the posterior predictive, which, using Jensen's inequality can also be seen to lower bound

$$\mathbb{E}_{\Pi_{\text{BEAR}}(p|X_{\text{train}})}[\text{KL}(p^{\infty}||p)],$$

where $\Pi_{\text{BEAR}}(p|X_{\text{train}})$ is the BEAR posterior learned from the training dataset.

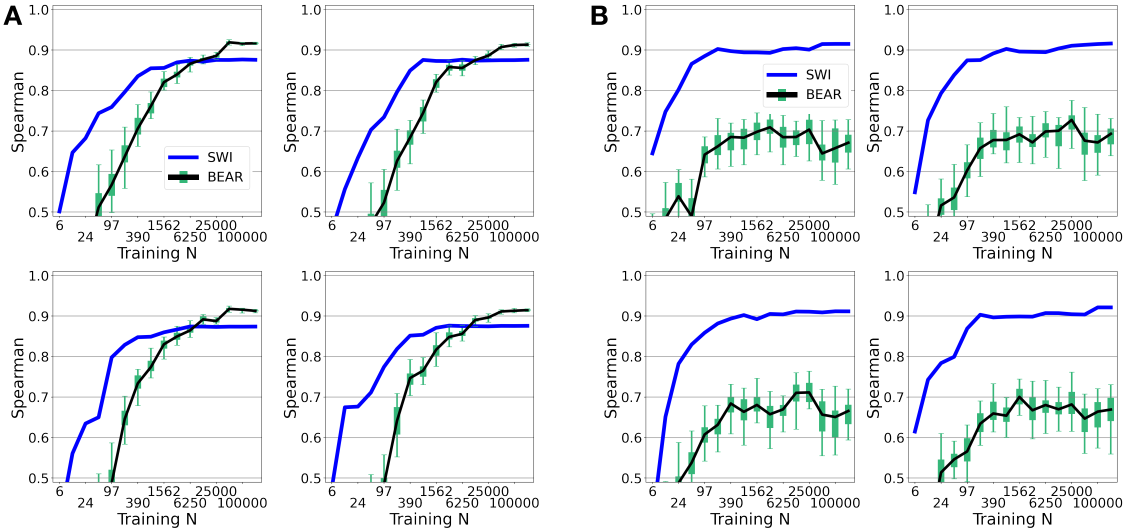


Figure S3: (A) Same as Fig. 4A, for four independent simulations following Scenario 1. (B) Same as Fig. 4B, for four independent simulations following Scenario 2.

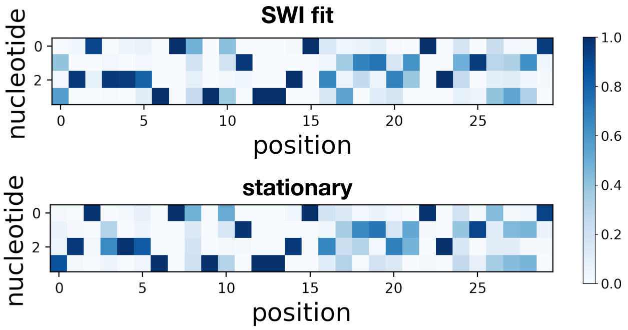


Figure S4: Probability of each nucleotide at each position learned by the SWI model (above) and in the stationary distribution p^{∞} (below), for a simulation from Scenario 2.

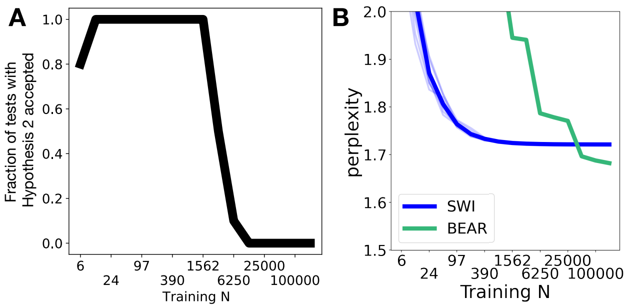


Figure S5: (A) Fraction of independent simulations (out of 10 total), following Scenario 1 (Sec. 6), in which Hypothesis 2 was accepted at level $\alpha = 0.025$. (B) Perplexity on heldout data of the BEAR and the SWI models in Scenario 1. Thick line corresponds to the average over 10 individual simulations (thin lines).

D Empirical Results Details

D.1 Data

Prediction task #1 (functional effect) Following standard practice, we report the absolute value of the Spearman correlation as $\mathcal{S}_f(p)$, since in some assays a negative change in the measured quantity corresponds to larger fitness (note that in all cases the predicted directionality of the effect under each model was correct). We focused on single amino acid substitutions, taking only those for which EVE was able to make a prediction (EVE is limited by its reliance on a multiple sequence alignment). We used the same data as in Shin et al. [2021], Table 1, taking the 37 experiments performed on the following 32 proteins: UBC9_HUMAN, UBE4B_MOUSE, P84126_THETH, HIS7_YEAST, BLAT_ECOLX, IF1_ECOLI, PTEN_HUMAN, B3VI55_LIPST, GAL4_YEAST, POLG_HCVJF, PABP_YEAST, CALM1_HUMAN, AMIE_PSEAE, TRPC_THEMA, RASH_HUMAN, YAP1_HUMAN, TRPC_SULSO, DLG4_RAT, BG_STRSQ, KKA2_KLEPN, HSP82_YEAST, B3VI55_LIPST (stabilized), MK01_HUMAN, HIV BF520 env, SUMO1_HUMAN, RL401_YEAST, PA_FLU, HG_FLU, TPMT_HUMAN, HIV BG505 env, TPK1_HUMAN, and MTH3_HAEAE (stabilized).

Prediction task #2 (pathogenicity) We used the pathogenicity labels of single amino acid substitutions curated from ClinVar [Landrum et al., 2018] in Frazer et al. [2021]. We considered labels for 87 human proteins less than 250 amino acids in length: AICDA, AQP2, ATPF2, B9D2, CAH5A, CAV3, CD40L, CF410, CHC10, CIA30, CLD16, CLN8, COQ4, CRBB2, CRGD, CTRC, CXB1, CXB2, CXB3, CXB4, CXB6, CY24A, DERM, DGUOK, DHDDS, EDAD, EFTS, ELNE, ETFB, ETHE1, EXOS3, FGF10, FGF23, FOXE3, FRDA, GP1BB, HBB, HEM4, HSPB1, HSPB8, IFM5, IFT27, JAGN1, KAD2, KCNE1, KCNE2, KITM, LITAF, MMAB, MMAC, MPU1, MYPR, NDP, NDUS8, NFU1, NKX25, NMNA1, OPA3, PAHX, PDYN, PMM2, PMP22, PNPH, PNPO, PROP1, PSPC, PTPS, RASH, RNH2A, S5A2, SAP3, SBDS, SCO1, SDHB, SDHF2, SIX1, SIX3, SOMA, TMM70, TNNT2, TPK1, TPM2, TR13B, TWST1, VHL, XLRS1, ZC4H2.

Training data All models were trained on datasets of protein sequences gathered as described in Shin et al. [2021] for pathogenicity effect prediction tasks and as described in Frazer et al. [2021] for functional effect prediction tasks. SWI and EVE were trained on the multiple sequence alignment, while Wavenet and BEAR were trained on the raw sequences as described in Shin et al. [2021]. All datasets were uniformly subsampled to produce a 75%/25% train/test split.

D.2 Models and code

The SWI model was trained via maximum likelihood.

The Wavenet model was trained via maximum likelihood with the default architecture, hyperparameters and training protocol described in Shin et al. [2021], for 100,000 steps. Code is from `https://github.com/debbiemarkslab/SeqDesign`. We did not apply the Wavenet model to the second prediction task, as it has only previously been developed for the first task.

The EVE model was trained via variational inference, using the same architecture, hyperparameters, and training protocol described in Frazer et al. [2021]. Code is from `https://github.com/debbiemarkslab/EVE`. To match the protocol of the original paper, EVE was – unlike SWI, Wavenet and BEAR – (a) trained on the full dataset rather than the training set alone, and (b) used a sequence reweighting heuristic.

The BEAR model used an embedded convolutional neural network (the same architecture as used in Amin et al. [2021], with layer 1 width of 16, filter width of 5 and 30 filters total) and a uniform prior over lags 2, 3, 5, 7, and 9. Code is from `https://github.com/debbiemarkslab/BEAR`. The model was trained using empirical Bayes, as described in Amin et al. [2021], for 500 steps with a batch size of 500000 kmers. To construct posterior credible intervals, we used 41 samples from the posterior for prediction task #1, and 1000 samples for prediction task #2.

We computed the heldout perplexity (Eqn. 26) for the BEAR posterior predictive and for Wavenet to produce Fig. S6.

D.3 Convergence experiments

To plot the convergence of the posterior over p_0 as a function of N (Fig. 5CD, S7 and S8), we used a vanilla BEAR model, a nonparametric Bayesian Markov model. Note that here we fixed the embedded AR model, rather than refitting with larger N, so that we could analyze the the convergence behavior with reference to

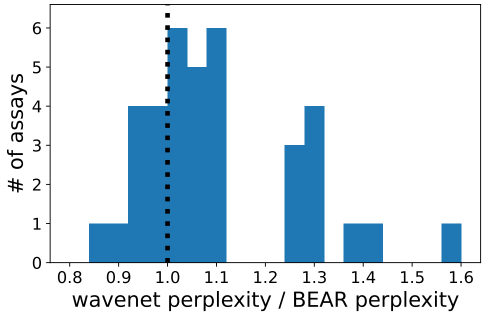


Figure S6: Ratio of the per residue perplexity on heldout data of the Wavenet model and of the BEAR model posterior predictive, across the 37 assays used for the first prediction task. Note lower perplexity corresponds to better density estimation performance.

the asymptotic results of Thm. 35 in Amin et al. [2021], which does not take into account empirical Bayes. We set the Dirichlet concentration to 10 and used a prior over lags as in Eqn. 25.

D.4 Interpolation experiments

We fit a BEAR model using the architecture and training protocol described in Sec. D.2, optimizing both the parameters of the AR model and h via empirical Bayes. We then varied h from its optimized value, and recalculated the total marginal likelihood and the posterior distribution over $\mathcal{S}_f(p)$ (Fig. 5EF and S9). We also computed the value of $\mathcal{S}_f(q_{\hat{\theta}})$ for the fit BEAR model in the $h \to 0$ limit (Fig. S10).

E Supplementary code

The supplementary code provides a Jupyter notebook (`example.ipynb`) illustrating the application of our BEAR diagnostic test on simulated data.

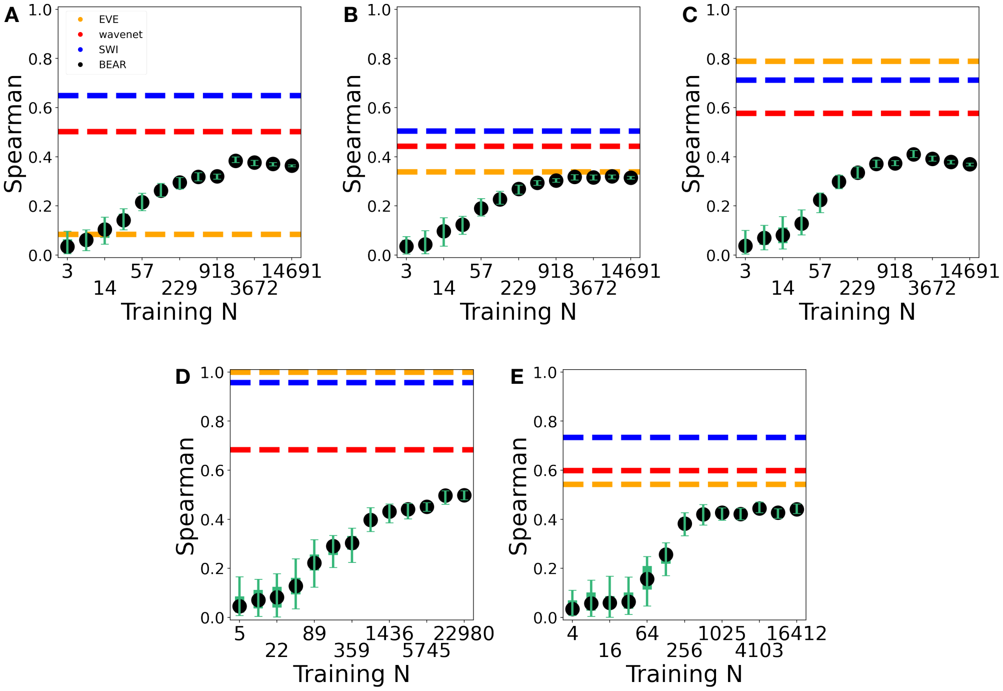


Figure S7: Same as Fig. 5CD, for 5 additional assay examples. A-C are each distinct β-lactamase assays; D is from GAL4 (DNA-binding domain); E is from UBE4B (U-box domain).

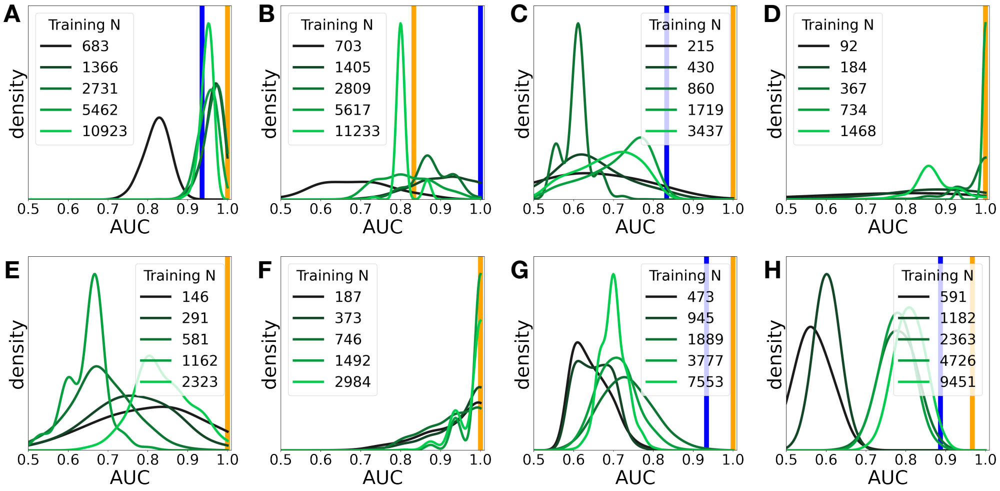


Figure S8: Convergence of the BEAR posterior over AUCs with N (green distributions), compared to the AUC of SWI (blue line) and EVE (yellow line), for the second prediction task. (A) is for the *CXB1* gene, (B) *CXB6*, (C) *EXOS3*, (D) *FGF23*, (E) *OPA3*, (F) *PAHX*, (G) *PROP1*, (H) *S5A2*.

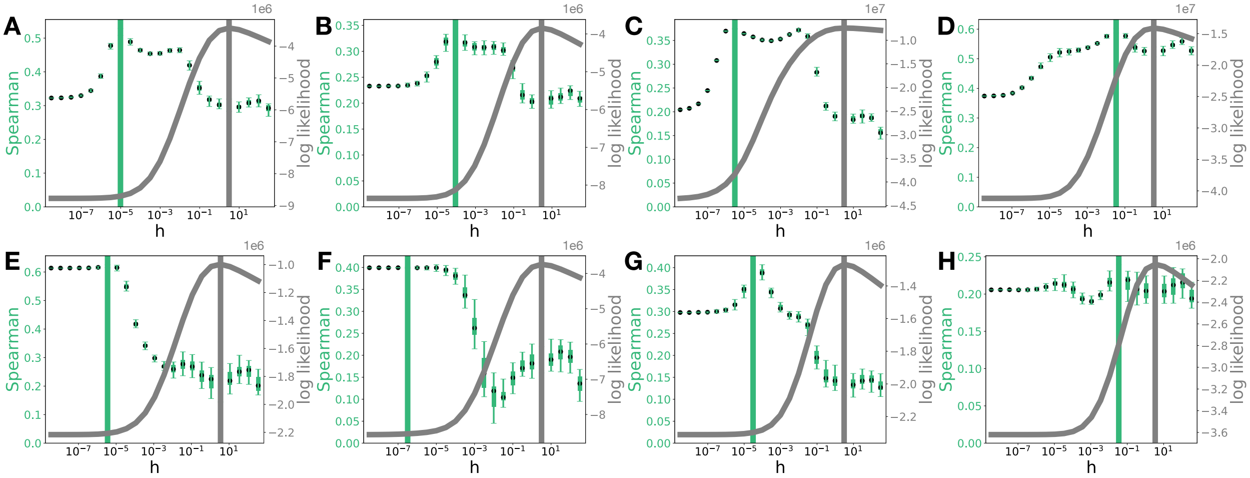


Figure S9: Same as Fig. 5EF, for 8 additional assay examples. (A) Aliphatic amidase, (B) levoglucosan kinase (stabilized), (C) HIV env protein (BF520), (D) β-glucosidase, (E) UBE4B (U-box domain) (F) TIM barrel, (G) thiopurine S-methyltransferase, (H) thiamin pyrophosphokinase 1.

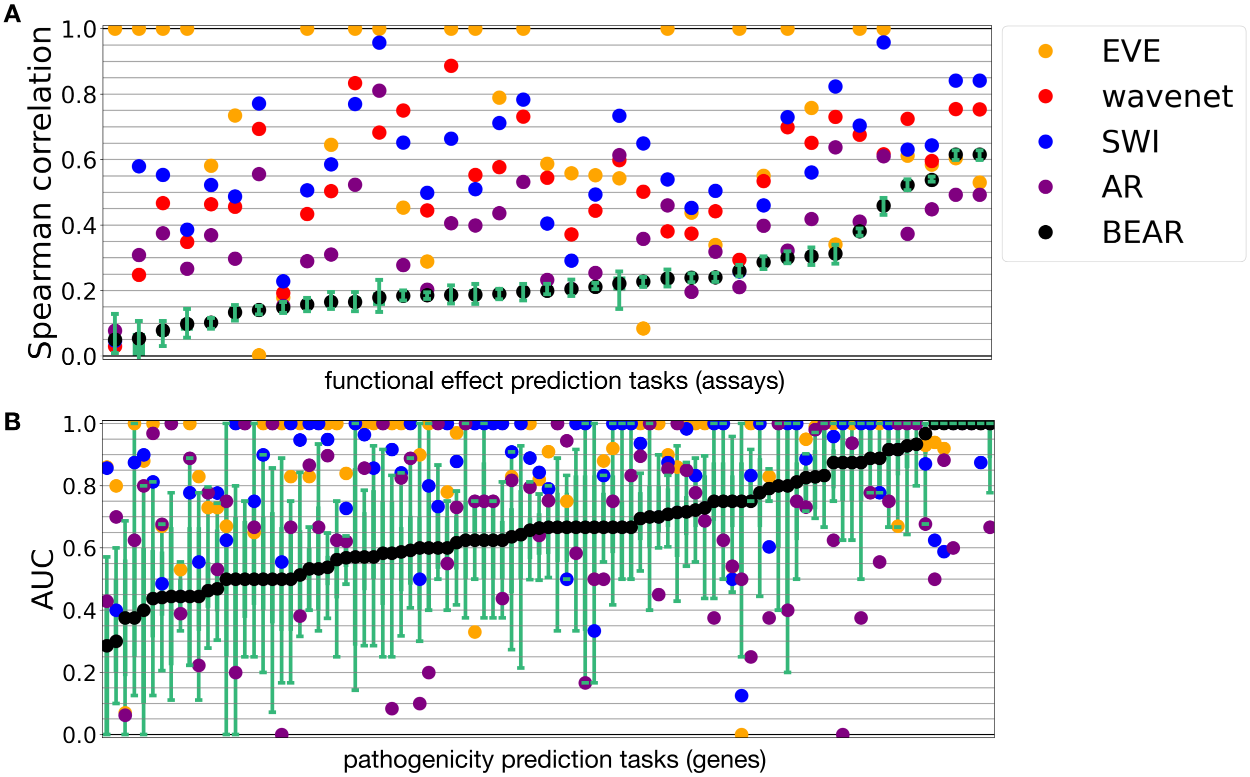


Figure S10: Same as Fig. 5AB, with the addition of the AR model in the $h \to 0$ limit (purple). In prediction task 1 (A), Hypothesis 2 is accepted in 28/37 assays (75%) while Hypothesis 1 is accepted in 6/37 (16%) for the AR model. In prediction task 2 (B), Hypothesis 2 is accepted in 16/97 genes (16%) and Hypothesis 1 is accepted in 17/97 genes (18%) for the AR model.