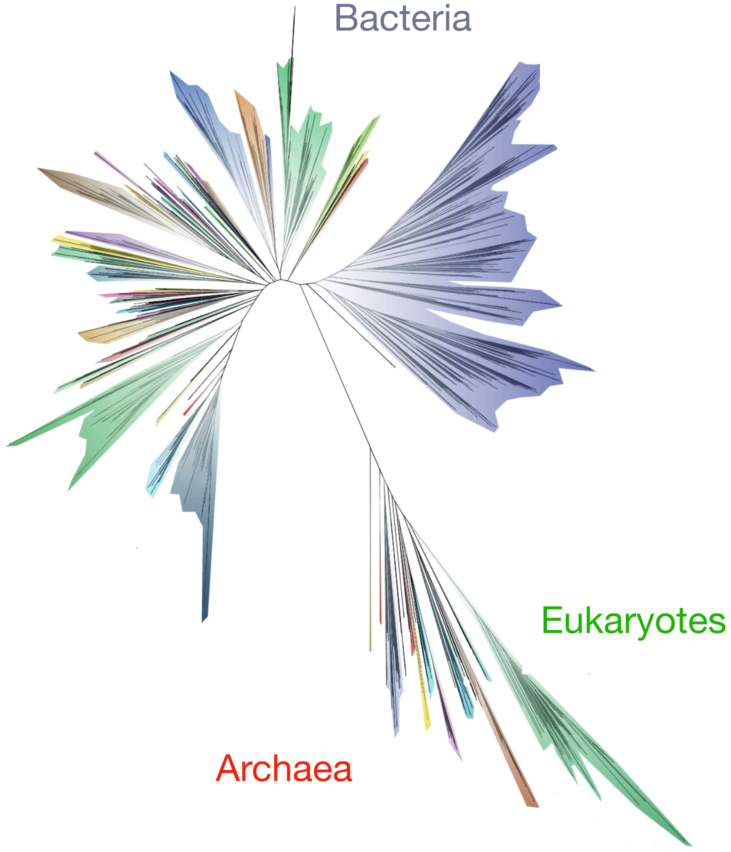
Non-identifiability & the Blessings of Misspecification in Models of Molecular Fitness

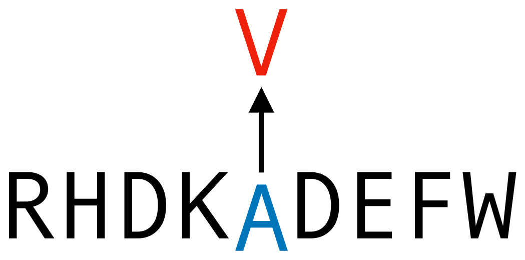
Eli N. Weinstein*, Alan N. Amin*, Jonathan Frazer, Debora S. Marks

October 11, 2022

Evolutionary data predicts **mutational effects**



Tree of life (Hug et al. 2016)

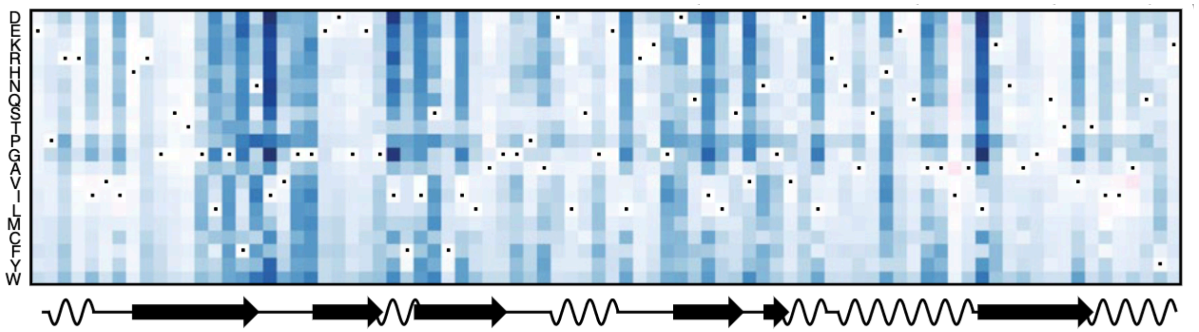


Long-term evolution of genome sequences

Laboratory measurements of molecular function

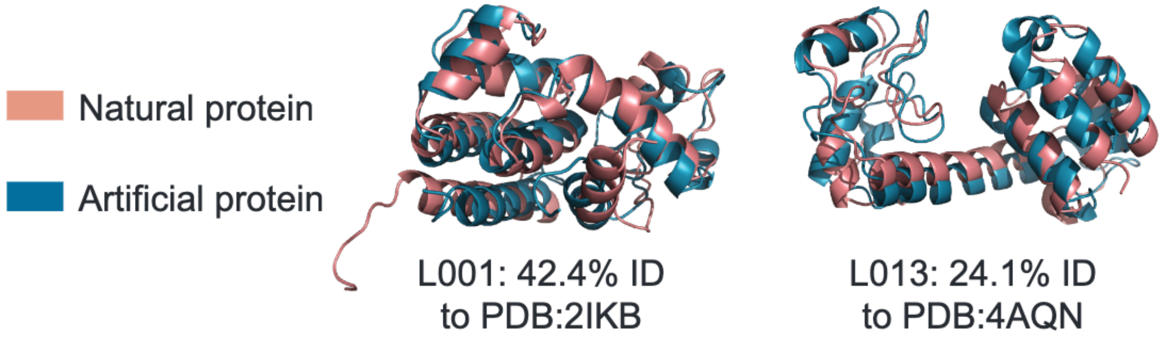
Applications

Hopf et al. 2017, *Riesselman et al. 2018*, Shin et al. 2021



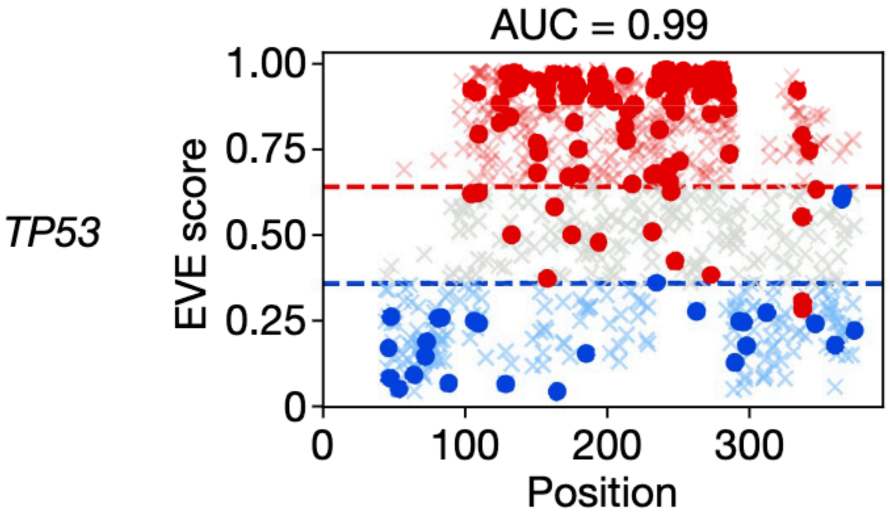
Predicting protein function

Madani et al. 2020, Russ et al. 2020, Shin et al. 2021



Designing new proteins

Hopf et al. 2017, *Frazer et al. 2021*



Predicting disease risk

Density Estimation and Fitness Estimation

Recipe for estimating molecular fitness from evolutionary data

1. Start with evolutionary sequence data, assumed to be iid

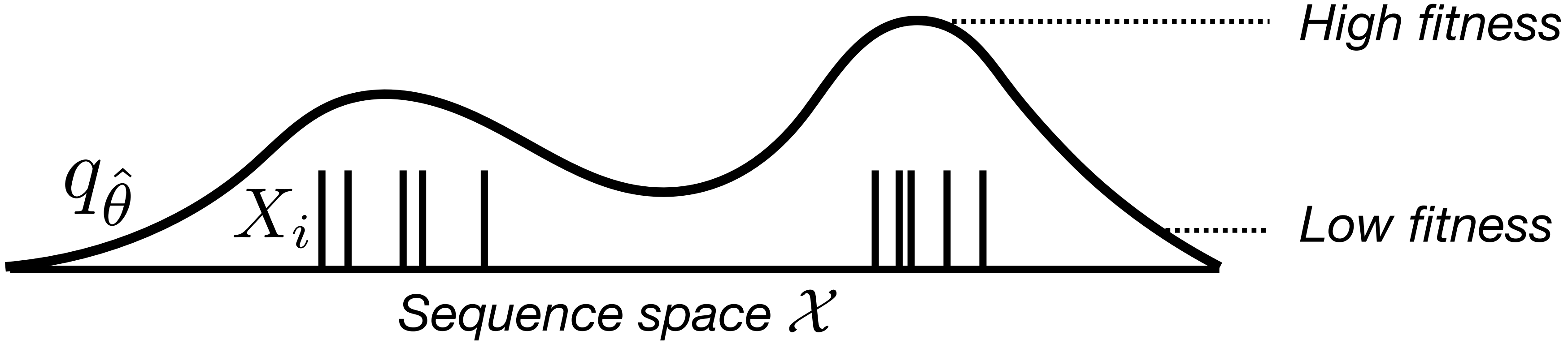
$$X_1, \ldots, X_N \sim p_0(x)$$

2. Fit a probabilistic model to the data

$$q_{\hat{\theta}} = \underset{q_\theta}{\operatorname{argmax}}\ q_\theta(X_{1:N})$$

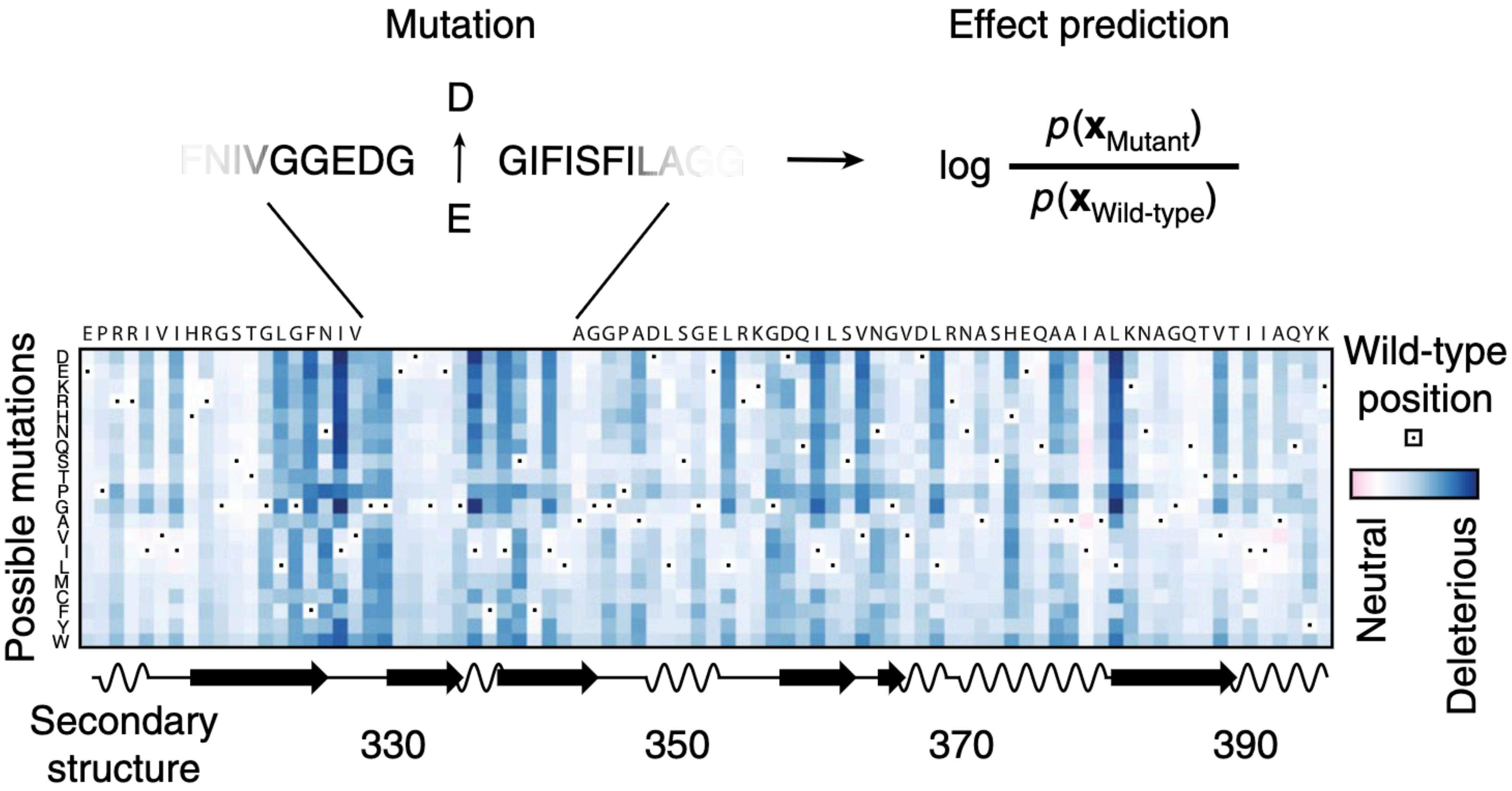
3. Use the inferred density as an estimate of fitness

$$\log q_{\hat{\theta}}(x) \approx \log p_0(x) \propto f(x)$$



Example

PDZ domain mutation effect predictions



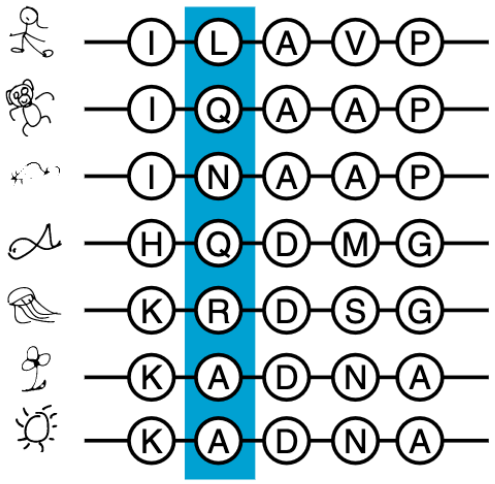
Progress in the field so far

Time

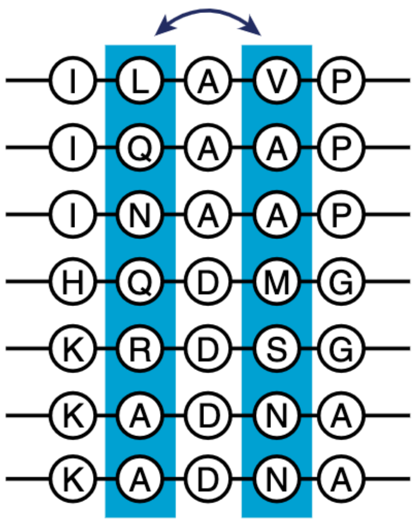
2017

2021

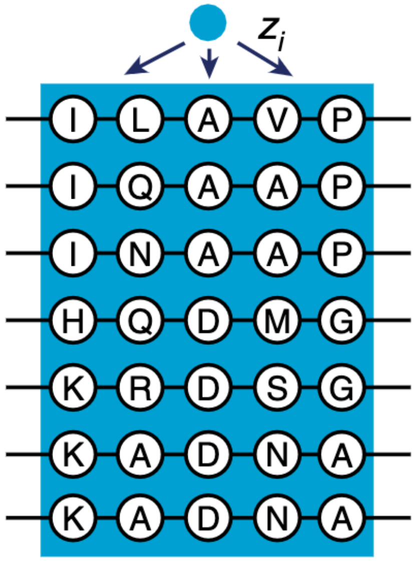
Site-wise independent models



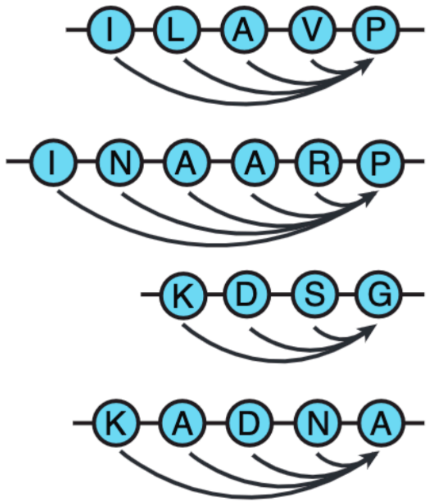
Potts models



Deep variational autoencoders



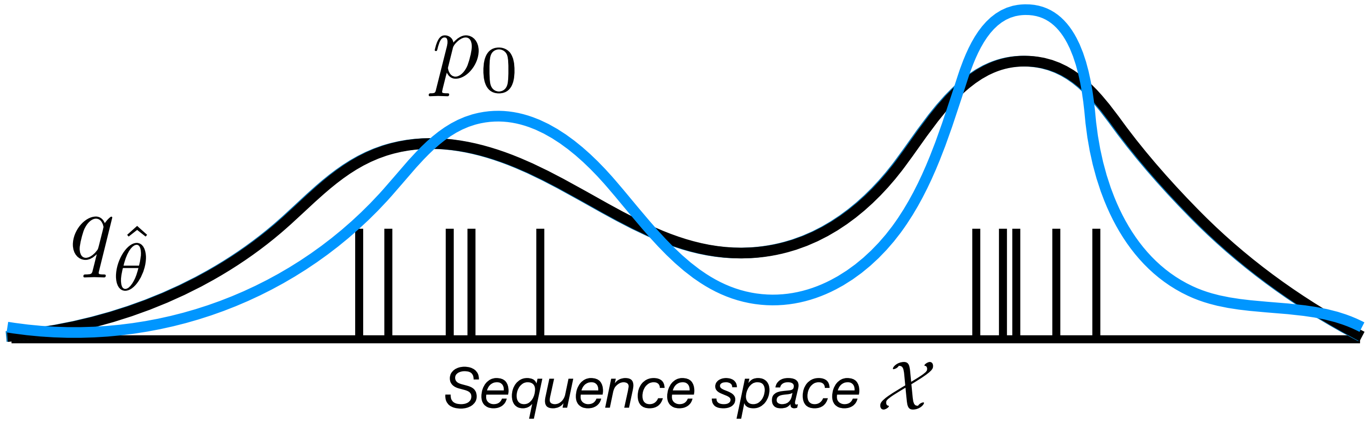
Alignment-free deep autoregressive models



Bigger models, bigger datasets

Naive hypothesis

Bigger, more flexible models → **Better density estimates** → **Better fitness estimates** **?**



Key Distinction

Data distribution

p_0 *True data distribution, i.e.* $X_1, X_2, \ldots \sim p_0(x)$

Target distribution

p^∞ *Reflects fitness* f*, i.e.* $p^\infty(x) = \frac{1}{\mathcal{Z}} \exp(\beta f(x))$

These two distributions may not be equal, for instance due to the effects of phylogeny.

Further, the target distribution in general is not identifiable given the data distribution.

Hypothesis #1: Misspecification is a Curse

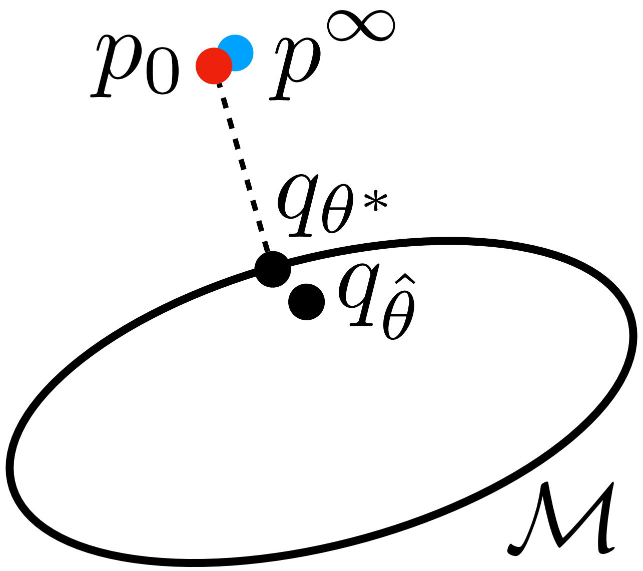
p_0 *True data distribution, i.e.* $X_1, X_2, \ldots \sim p_0(x)$

p^∞ *Reflects fitness, i.e.* $p^\infty(x) = \frac{1}{\mathcal{Z}} \exp(\beta f(x))$

$q_{\hat{\theta}}$ *Model fit to observed data*

Hypothesis #1

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



Better density estimation = better fitness estimation

Hypothesis #2: Misspecification is a Blessing

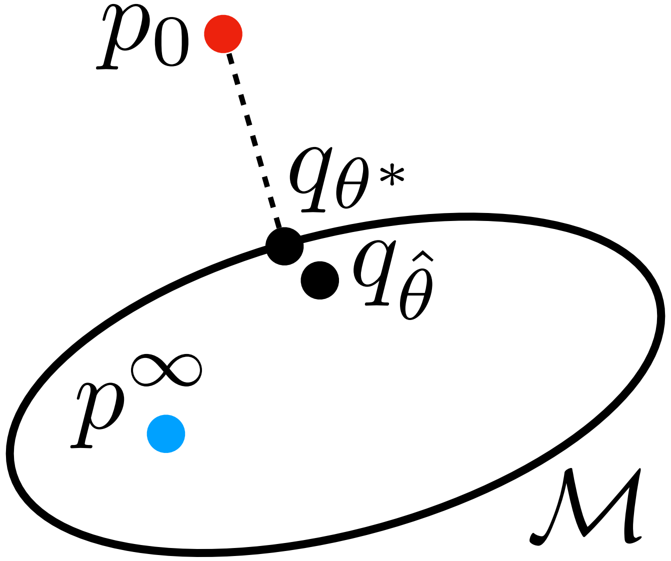
p_0 *True data distribution, i.e.* $X_1, X_2, \ldots \sim p_0(x)$

p^∞ *Reflects fitness, i.e.* $p^\infty(x) = \frac{1}{\mathcal{Z}} \exp(\beta f(x))$

$q_{\hat{\theta}}$ *Model fit to observed data*

Hypothesis #2

Fitness estimation methods succeed by using models $\mathcal{M}$ that are misspecified with respect to p_0, i.e. $p_0 \notin \mathcal{M}$. Then $q_{\hat{\theta}}$ is then closer to p^∞ than p_0.



Worse density estimation = better fitness estimation

When Could Misspecification Help?

Large data limit of model: $q_{\theta^*} = \operatorname{argmin}_{q_\theta \in \mathcal{M}} \mathrm{KL}(p_0 \| q_\theta)$

Log-convex model: For any $\theta, \theta' \in \Theta$ and $0 < r < 1$, there exists some θ'' such that $q_{\theta''}(x) = q_\theta(x)^r q_{\theta'}(x)^{1-r} / \sum_x q_\theta(x)^r q_{\theta'}(x)^{1-r}$

Theorem:

Assume that the model $\mathcal{M}$ is log-convex and $p^\infty \in \mathcal{M}$. Then, if $p_0 \notin \mathcal{M}$,

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

But if $p_0 \in \mathcal{M}$,

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

Progress in fitness estimation:

1. **Hypothesize** models where $p^\infty \in \mathcal{M}$ and $p_0 \notin \mathcal{M}$
2. **Check** predictions against experimental fitness measurements.
3. **Iterate.**

Key Tool: Nonparametric Density Estimator

Bayesian Embedded Autoregressive (BEAR) Model

Amin, Weinstein* & Marks, NeurIPS 2021*

Theorem (Posterior consistency):

For $M > 0$ sufficiently large and $\epsilon \in (0, 1/2)$ sufficiently small,

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

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

Unbiased: converges to *any* p_0, no matter how complicated.

Quantifies uncertainty: gives range of possible p_0 compatible with the data.

Diagnostic Test

$\mathcal{S}_f(p)$ Score evaluating how accurately p predicts fitness f based on external experimental/clinical data.

Diagnostic test

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*

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

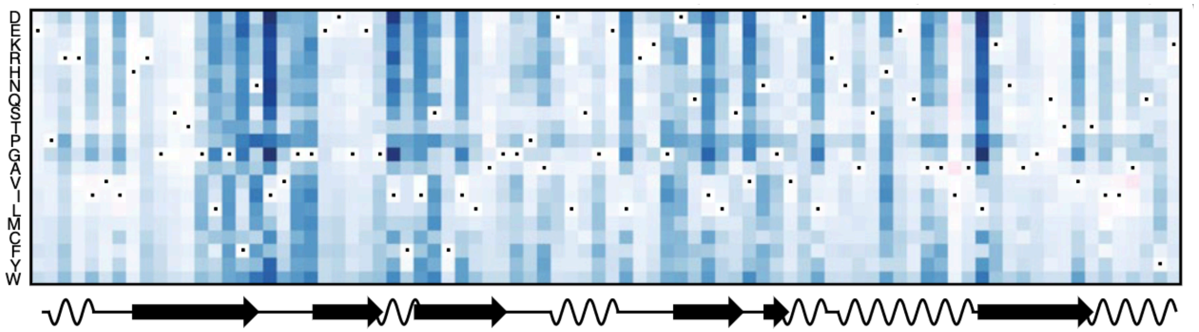
Accept Hypothesis 1 at significance level α *if*

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

Fitness Prediction Tasks

1. Experimental assays of protein function

Evaluate with Spearman correlation between assay output and log probability.
37 assays, 32 protein families, ~1000s of measurements per assay.

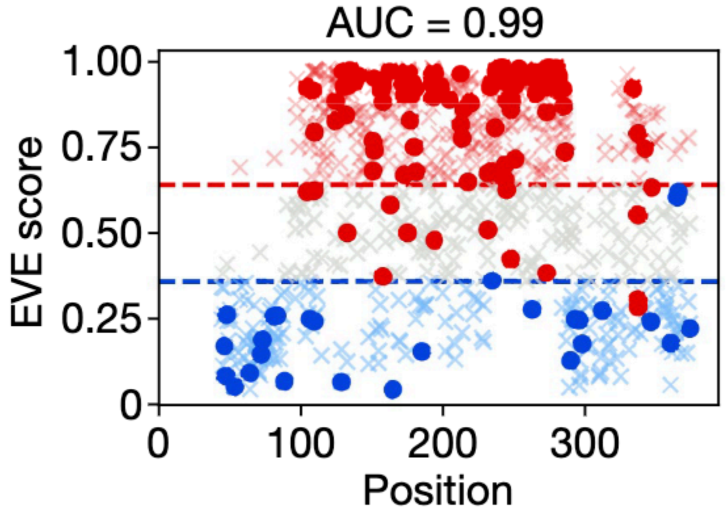


Hopf et al. 2017,
Riesselman et al. 2018,
Shin et al. 2021

2. Clinical annotation of variant disease risk

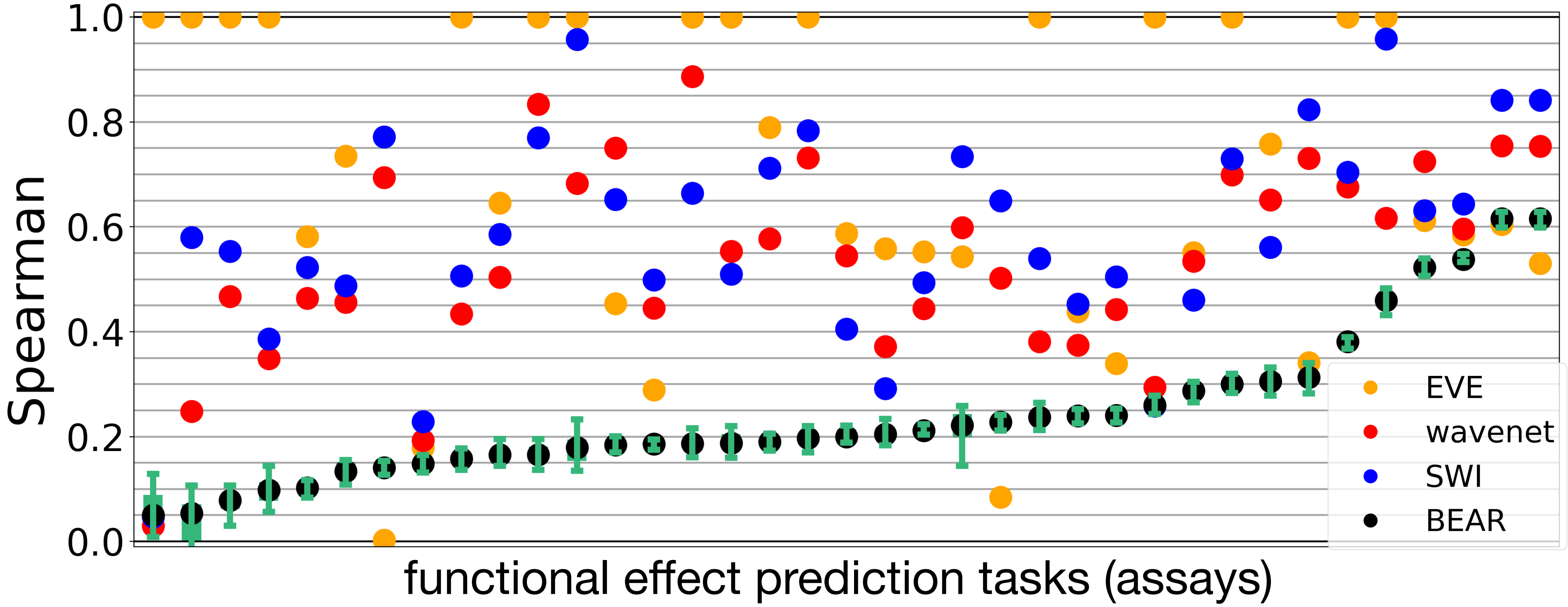
Evaluate with AUC when log probability is used to predict variant pathogenicity
97 genes, 87 protein families, ~1-10 measurements per assay.

TP53



Hopf et al. 2017,
Frazer et al. 2021

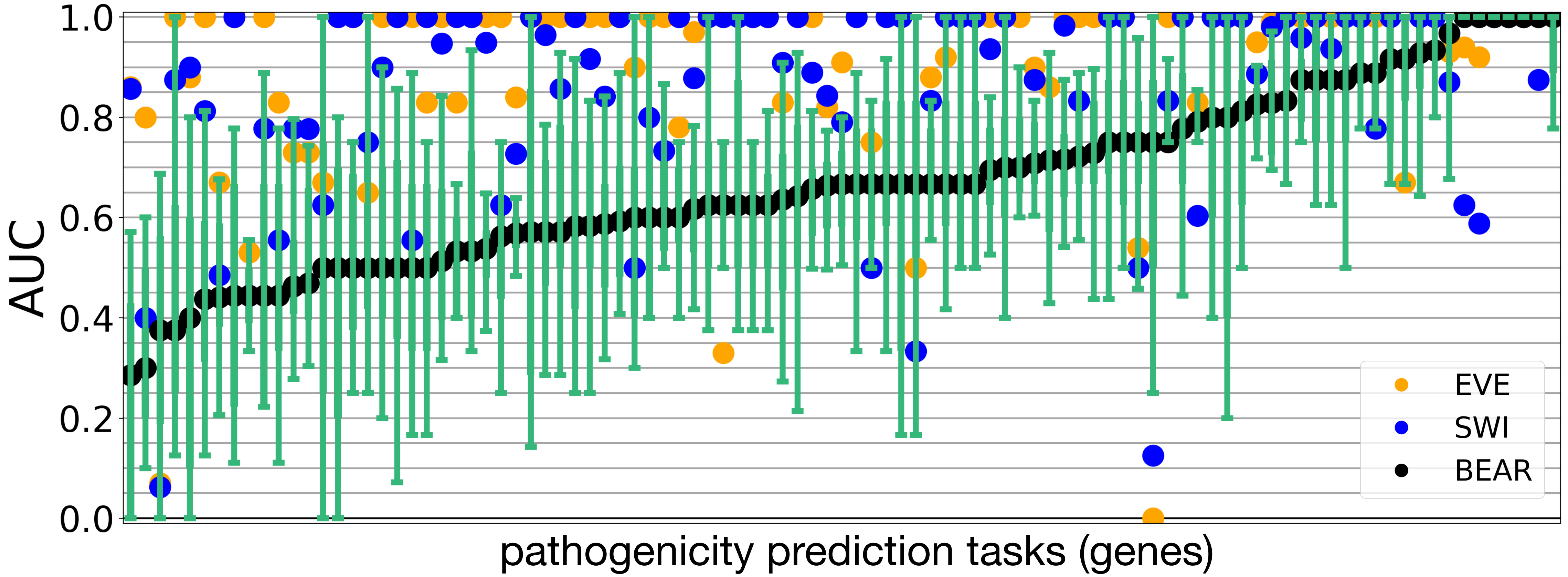
Results: Experimental Assays



Hypothesis 2 is strongly preferred.

Existing models systematically outperform the true data distribution.

Results: Clinical Disease Risk



Hypothesis 2 is strongly preferred.

Existing models systematically outperform the true data distribution.

Conclusions

- Fitness and phylogeny are non-identifiable.
- Better density estimation can lead to worse fitness estimation.
- Existing fitness estimation methods succeed because of, not despite, misspecification.
- **Progress through bigger models, trained on bigger datasets, is not inevitable.**

Non-identifiability & the Blessings of Misspecification in Models of Molecular Fitness

Eli N. Weinstein*, Alan N. Amin*, Jonathan Frazer, Debora S. Marks