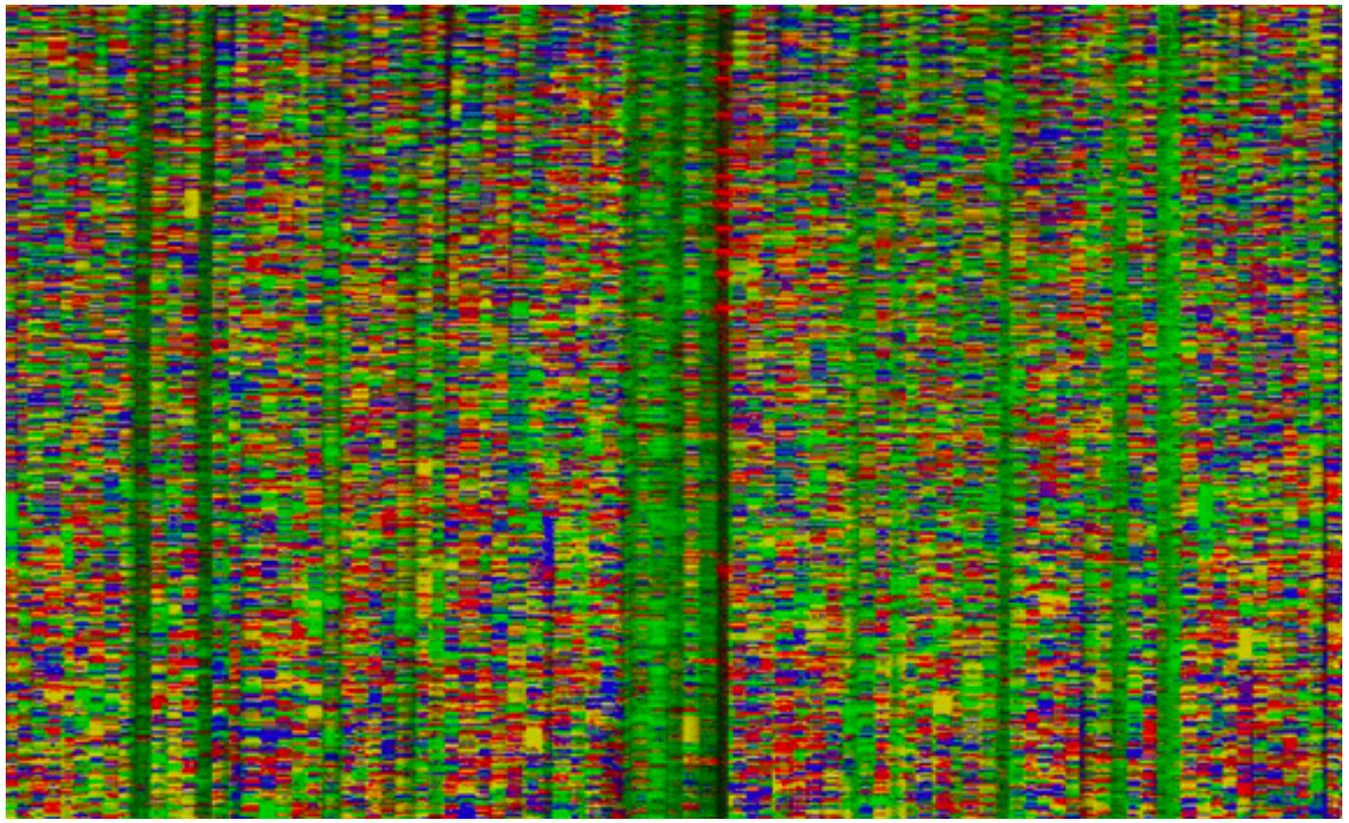
A structured observation distribution for generative biological sequence prediction and forecasting

Eli N. Weinstein and Debora S. Marks

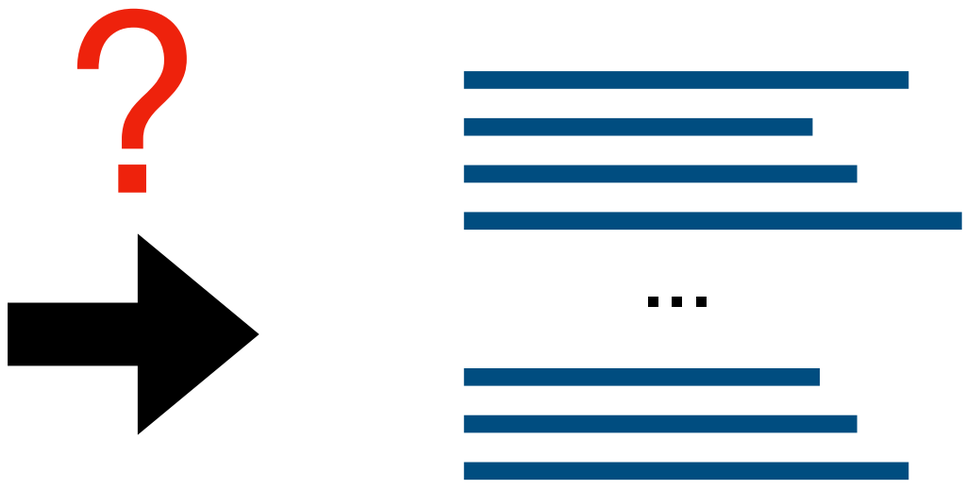


Human genome project-era fluorescent sequencer readout,
Sanger Institute, Wellcome Image no. B0002668

- **Measuring** and **making** sequences is central to modern biology.
- Evolutionary biology, immunology, oncology, microbiology, therapeutics, ...
- This talk is about using probabilistic machine learning to analyze, predict and generate sequences

Designing models: from vectors to sequences

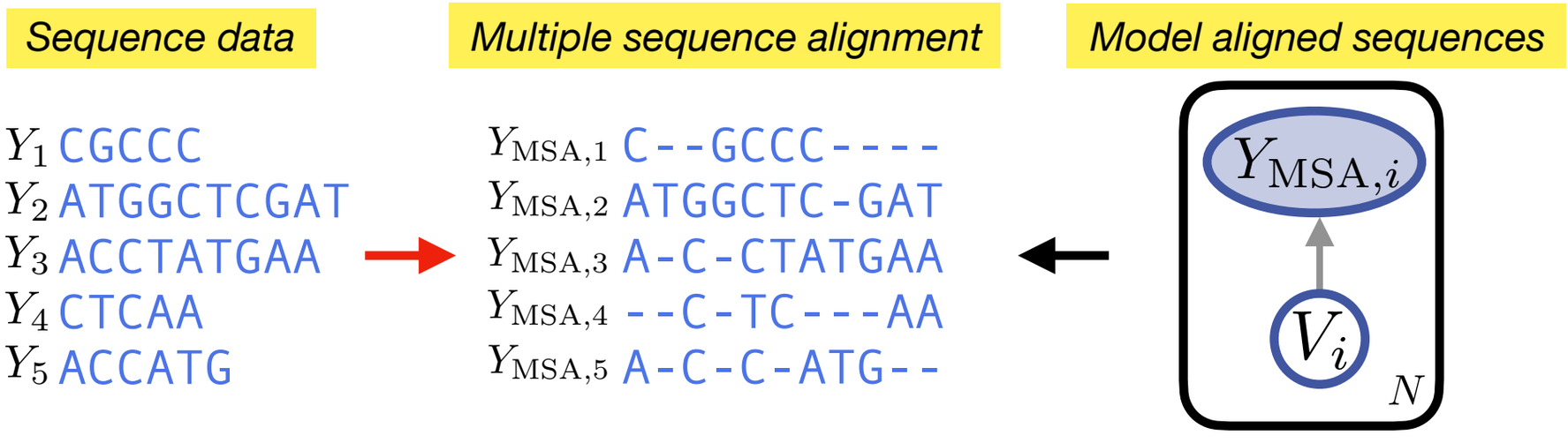
Linear regression, Gaussian processes, principal component analysis, independent component analysis, ordinary differential equations, stochastic differential equations, variational autoencoders, etc.



biological sequences (individual proteins, RNA, etc.)

- Models of continuous vectors or matrices are ubiquitous and useful.
- We want to apply these models to biological sequences.
- Problem: data lives in a different space with a different notion of distance.

Conventional approach

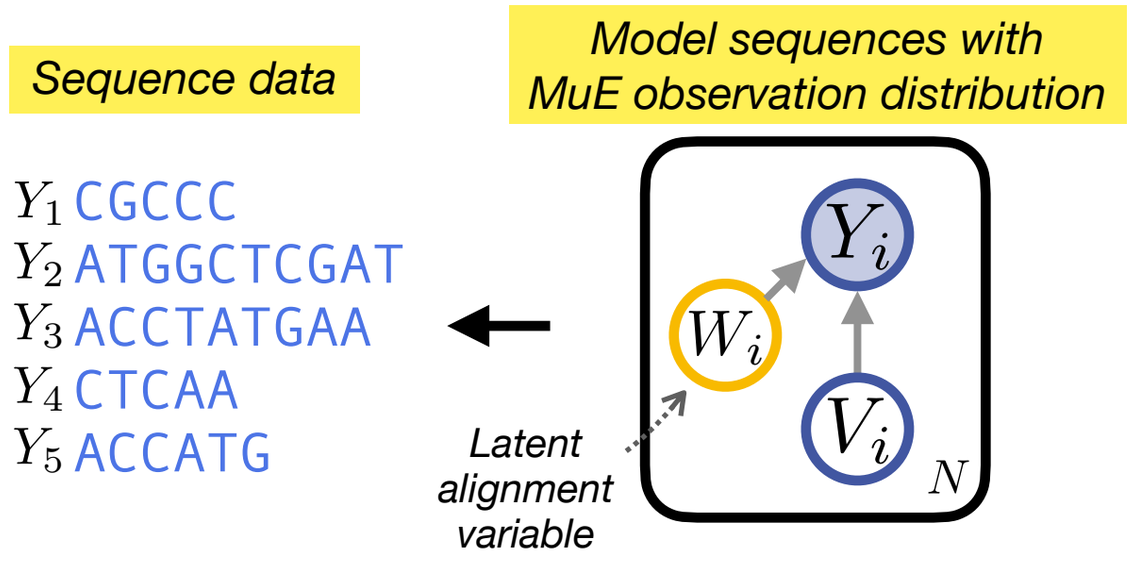


$$\text{Preprocess: } \{Y_{\text{MSA},1}, \ldots Y_{\text{MSA},N}\} := f_{\text{MSA}}(\{Y_1, \ldots, Y_N\}),$$
$$\text{Model: } V_i \sim p_\theta(v) \quad Y_{\text{MSA},i} \sim \text{Categorical}(\text{softmax}(V_i)).$$

- MSA captures **fundamental biology**: there are conserved positions across similar sequences, and mutations are mainly substitutions and indels.
- Building models this way **violates fundamental statistical assumptions**: past data changes as more data is added, data dimension (the space data lives in) changes as more data is added.

➡ *i.i.d. assumptions break down, can't evaluate sequence predictions/forecasts*

Proposed approach



$$\textbf{Model: } V_i \sim p_\theta(v) \quad Y_i \sim \mathrm{MuE}(\mathrm{softmax}(V_i), c, \ell, a^{(0)}, a^{(t)})$$

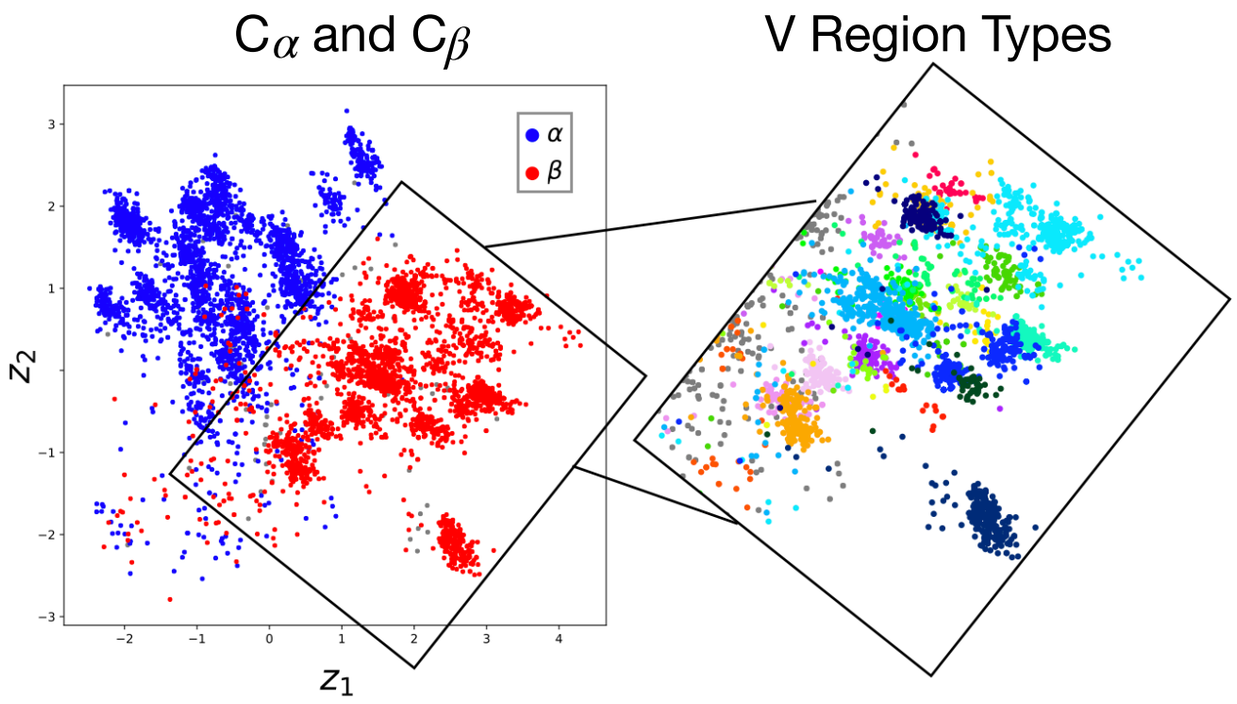
- Replace preprocessing step with generative process. Instead of filtering indels out of the data, add them in the model.
- Extend continuous vector model with our new **mutational emission (MuE)** distribution.
- Model retains the key ideas behind alignment: can still talk about variation at conserved sites, indels, etc..

Immune receptor repertoires

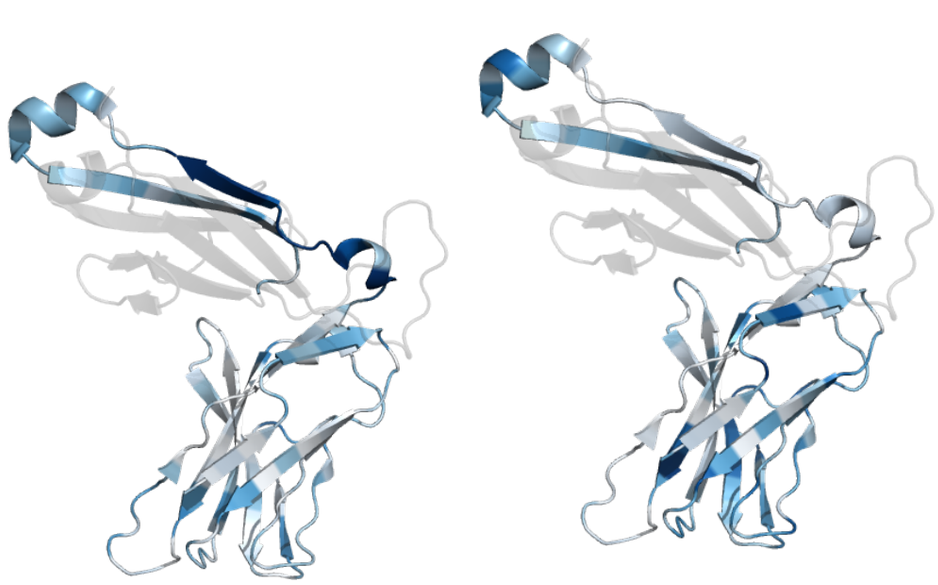
Improved predictive performance

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

Informative latent representations

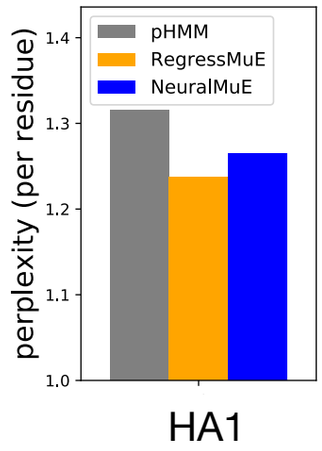


Features at conserved sites

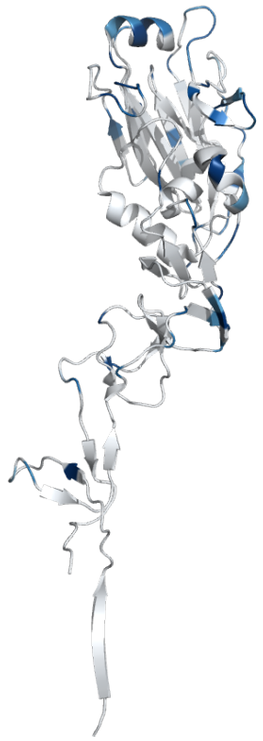


Generative forecasts of viral evolution

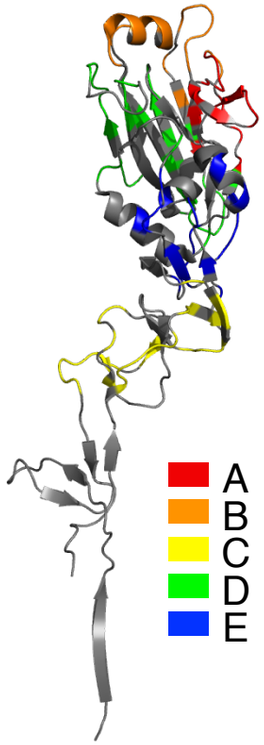
Improved predictive performance



Informative regression coefficients

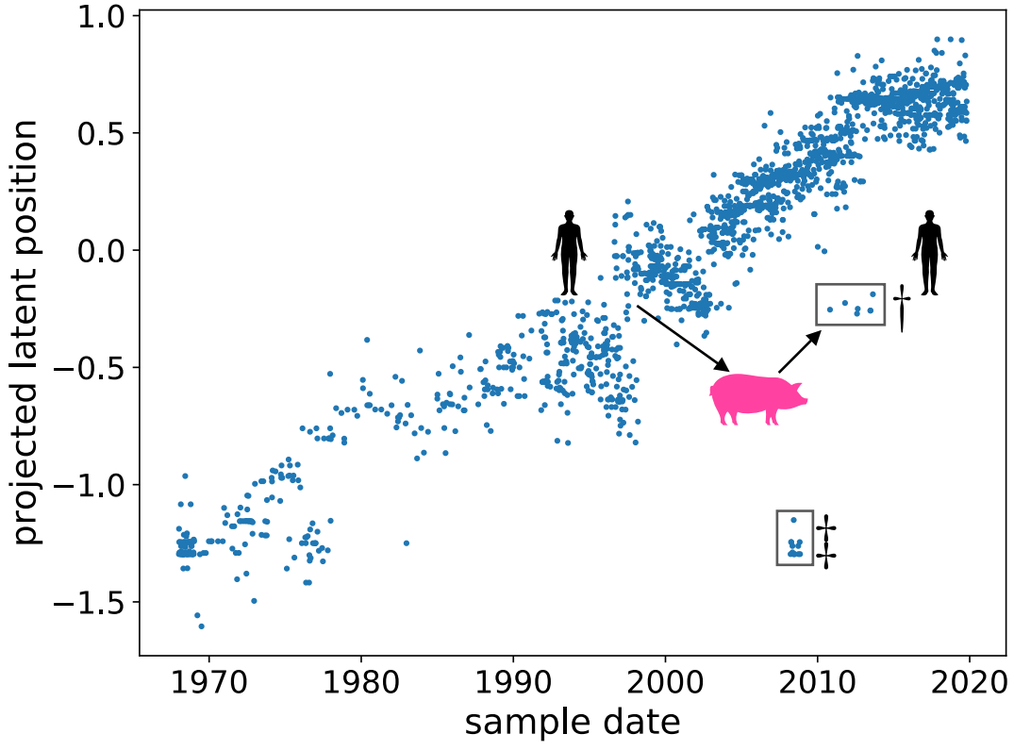


Inferred shift ν, 1968-2019



Epitope regions

Informative latent representations



Conclusions

- The MuE enables application and rigorous evaluation of a wide range of statistical models (vector models, including factor models, regression models, image models, etc.) to biological sequences.

- The MuE both accounts for common forms of variation and avoids the serious pathologies of MSA preprocessing, a ubiquitous technique used to account for the same variation.

- The MuE is now part of the Pyro probabilistic programming language, providing a platform for building new models and integrating information across data sources.