

Phylogenetic corrections and higher-order sequence statistics in protein families: Potts vs multiple sequence alignment transformer machine learning models

Kisan Khatri¹, Ronald M. Levy², and Allan Haldane^{1,*}

¹Department of Physics and Center for Biophysics and Computational Biology, Temple University, Philadelphia, Pennsylvania 19122, USA

²Department of Chemistry and Center for Biophysics and Computational Biology, Temple University, Philadelphia, Pennsylvania 19122, USA



(Received 28 February 2025; accepted 7 September 2025; published 20 October 2025)

Recent generative machine learning models applied to protein multiple sequence alignment (MSA) datasets include simple and interpretable physics-based Potts covariation models and other machine learning models such as MSA Transformer (MSA-T). The best models accurately reproduce MSA statistics induced by the biophysical constraints within proteins, raising the question of which functional forms best model the underlying physics. The Potts model is usually specified by an effective potential including pairwise residue-residue interaction terms, but it has been suggested that MSA-T can capture the effects induced by effective potentials that include more than pairwise interactions and implicitly account for phylogenetic structure in the MSA. Here, we compare the ability of the Potts model and MSA-T to reconstruct higher-order sequence statistics reflecting complex biophysical sequence constraints. We show how the model performance depends greatly on the treatment of phylogenetic relationships between the sequences, which can induce nonbiophysical mutational covariation in MSAs. When using explicit corrections for phylogenetic dependencies, we find that the Potts model outperforms MSA-T in detecting epistatic interactions of biophysical origin. Furthermore, the sequences generated by the models we studied here are predicted to fold into nativelike structures, as indicated by AlphaFold predictions.

DOI: [10.1103/g5cx-1vhj](https://doi.org/10.1103/g5cx-1vhj)

I. INTRODUCTION

Machine learning models have made great strides in predicting the functional and structural properties of proteins based on large protein sequence datasets, including the subclass of generative protein sequence models (GPSMs) that learn an underlying sequence probability distribution $P(S)$ from a multiple sequence alignment (MSA) to generate new synthetic protein sequences S . To function well, GPSMs must capture covarying amino acid patterns in the MSA that implicitly encode information about physical constraints on proteins, enabling the design and detection of hidden protein properties from sequence data [1,2]. This raises the question of which GPSM functional form best describes protein biophysics, which best distinguishes the different causes of covariation, and how to measure this.

Mutational covariation in protein sequences provides a highly informative statistical signal that accurate GPSMs must capture. Covariation arises from multiple factors, including the following [3]: (1) High-fitness sequence motifs and epistasis (mutational cooperativity) underlying biophysical function lead to compensatory mutation pairing patterns between residues. (2) Phylogenetic relationships due to recent speciation or gene duplication can distort covariation patterns, causing “excess” counts of patterns from

evolutionarily related sequence clusters as illustrated in Fig. 1. Statistical noise due to the limited number of available distinct natural sequences used during model inference introduces finite-sampling statistical variation in estimated covariation and in GPSM accuracy. The latter two can be considered nuisance factors in uncovering the underlying biophysics driving the first factor of fitness-induced covariation. These covariation signals are measured from MSAs of individual protein families, which consist of homologous proteins with shared function and overall structure.

We examine how two leading GPSMs in recent focus distinguish these effects, the Potts model [1,4–9] and the MSA Transformer (MSA-T) [10,11], with careful control for how phylogenetic clusters affect GPSM performance. We suggest that the most reliable metrics to measure GPSM performance use statistical characteristics of individual protein families designed to exclude the biasing effects of phylogenetic relationships between sequences that do not originate from the fundamental biophysical properties of the protein family. We find that when the impact of phylogenetic relationships is accounted for, the Potts model better captures one such set of statistics, the higher-order MSA statistics, referring to the covariation patterns involving more than two amino acids. We expect these complex mutational dependencies to reflect epistatic networks underlying protein function, structure, and evolution [12], forming a more challenging test than capturing simple pairwise covariation.

II. GPSM ARCHITECTURES

Despite their quite different architectures and complexity, the Potts model and MSA-T exhibit fundamental

*Contact author: allan.haldane@temple.edu

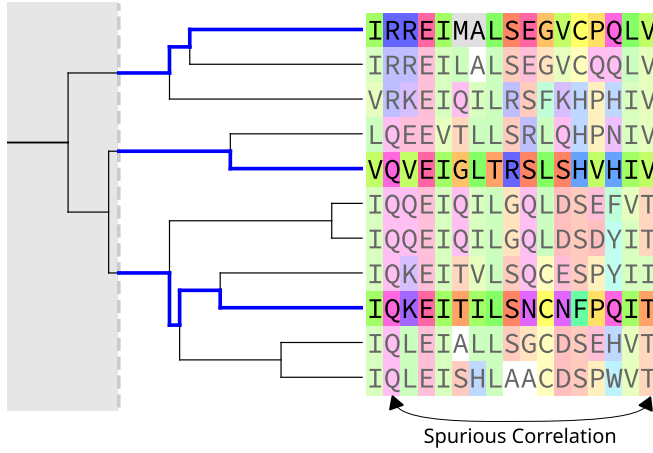


FIG. 1. Phylogenetic relationships between sequences in an MSA result in a spurious mutational correlation due to common ancestry, for R/V and Q/T combinations at the illustrated column pair. Sequence pairs greater than an identity cutoff (diverging to the left of the gray dotted line) approximate i.i.d. samples, so that identity filtering to retain three sequences labeled in blue gives an unbiased sample showing no correlation with equal frequency of R/V, Q/V, and Q/T combinations.

commonalities in their design to account for covariation patterns [13]. The Potts model is a statistical physics model inspired by spin glasses that is increasingly being adapted as a generative modeling tool in machine learning to capture sequence variability within a single protein family. A Potts model trained on protein family F has probability distribution $P(S | \theta_F) \propto \exp(-\sum_{i < j} J_{s_i s_j}^{ij})$ for sequences S to evolve, with characters s_i at position i , and pairwise “coupling” parameters $\theta_F = \{J_{\alpha\beta}^{ij}\}$ measuring the favorability of having residues α, β at positions i, j in a sequence. It models complex higher-order correlations through networks of pairwise interactions. In contrast, MSA-T is a deep learning [14] masked language neural network attention model [15] trained only once on MSAs of all available protein families simultaneously [10] with about 10^4 times more parameters, recently used for protein structure and evolution analyses [16–22]. The Potts model has been found to have superior generative accuracy to some other GPSMs, including variational autoencoders and site-independent models [12], but a generative method subsequently developed for MSA-T has been reported to better reproduce higher-order statistics [23], though its parameters lack clear physical interpretation and the reasons for this result are ambiguous.

III. EFFECT OF PHYLOGENETIC RELATIONSHIPS ON GPSM EVALUATION

GPSMs must distinguish the causes of covariation, and in the case of the Potts model, an identity filtering procedure is critical. The Potts model is trained to reproduce the site statistics of a training MSA from the protein family, assuming each sequence is an independent and identically distributed (i.i.d.) sample. To be i.i.d., the sequences can be envisioned to have evolved from a distant ancestral sequence under a common fitness function and mutational process

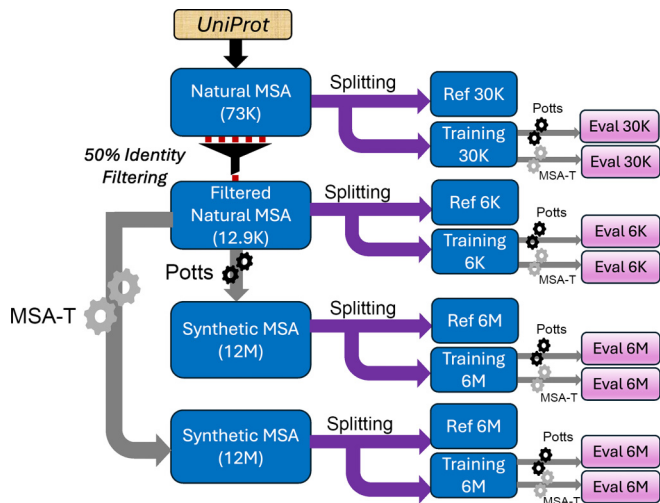
encoded in $P(S | \theta_F)$, and enough time must have passed for any statistical correlation with each other to be effectively nil due to mutational saturation. Phylogenetic relationships violate the *independence* assumption, as clusters of orthologs and other closely evolutionarily related sequences are explicitly non-i.i.d. In standard Potts methodology, the related sequences are filtered to eliminate phylogenetic redundancy.

In contrast, MSA-T accounts for phylogenetic relationships through a columnwise attention layer [11,21], and is not explicitly trained on MSA statistics but rather on a masked prediction task insensitive to phylogenetic structure, in predicting character states in input MSAs from all protein families at once, which were randomly masked. A method has been proposed to use MSA-T in a generative fashion for a single protein family by repeatedly masking and resampling an input MSA of that family to produce novel sequences [23]. The MSA-T probability distribution for this generative method is $P(S | \theta, M_F, p)$ for generating sequence S , and depends the model’s pretrained parameters θ , an input MSA M_F for a protein family F , and a masking frequency parameter p . Unlike the Potts model that is trained separately for each protein family, MSA-T is trained only once on all protein families, and the generation of new sequences does not involve any additional training but instead iterated prediction given an initial MSA of a given family.

One goal of this study is to control for how phylogenetic clusters affect GPSM performance. After filtering, the i.i.d. MSA will have covariation signals that exclude biases due to clustering and so are presumably biophysical in origin, and should be reproducible in arbitrary subsets of the MSA. This provides a foundation for determining which GPSMs best capture biophysical constraints.

IV. STATISTICAL METHODOLOGY

We perform tests of GPSM performance with and without accounting for phylogeny as outlined in Fig. 2. We generate sequences from MSA-T using the default method of Ref. [23] using precomputed parameters θ , and train the Potts model using the high-accuracy Mi3-GPU method [24]. We also test a very simple GPSM, the Independent model, with distribution $P(S) = \prod_i f_{s_i}^i$, where f_{α}^i are single-site frequencies for amino acid α at position i found in the training MSA, which is unable to accurately capture even the pairwise amino-acid frequencies $f_{\alpha\beta}^{ij}$ of the training MSA. Given various training MSA data described in subsequent sections, we generate new MSA datasets from each GPSM and then evaluate GPSM performance by comparison of the generated “evaluation” MSA to a “reference” MSA using a metric r_{20} , which measures higher-order covariation and provides a more stringent test of GPSM accuracy than pairwise covariation statistics or point-mutant effects [12]. The training, reference, and evaluation MSAs have the same number of sequences. These tests are designed to isolate the effects of fitness, phylogeny, and statistical noise, and to control statistical errors including model specification, out-of-sample, and estimation errors [12], such as by splitting each MSA dataset into training and reference MSAs to ensure that no sequences used to train the model are used in its evaluation. We examine two protein families: RR domain (PF00072), which was previously studied to test



To characterize these patterns, we compute the frequencies of noncontiguous amino acid “words” of length n , or higher-order marginals (HOMs) in both reference and model-generated evaluation MSAs, as in Ref. [12]. For each HOM length, we randomly select 3000 position sets, extract the top 20 most frequent words, and calculate the Pearson correlation (r) between their frequencies for each position set in those two MSAs. The averaged r over all sets is defined as r_{20} score, a metric that evaluates the model’s ability to reproduce higher-order sequence statistics and model complex chains and networks of interactions within protein [12]. This statistics for $n = 2$ measures how well the model reproduces the pairwise site statistics, and in the Supplemental Material [25], we illustrate the model performance for predicting pairwise mutational covariation (Fig. S1). As a reference point, we also compute a “null” result that represents the maximum attainable value for the r_{20} metric for a well-specified model subject only to finite-sampling limitations in model evaluation, in other words for an idealized GPSM with no specification or out-of-sample error. It is computed by measuring r_{20} for two independent MSAs drawn from the same probability distribution, the training MSA and reference MSA [12]. This provides a baseline for quantifying how accurately GPSMs model the target distribution, and deviations from the null result indicate imperfect modeling.

Figure 2 consists of two line plots, (a) and (b), showing the relationship between r_{20} (Y-axis, ranging from 0.0 to 1.0) and Pattern Length (X-axis, ranging from 2 to 8). The legend indicates four models: Potts (blue circles), MSA-T (orange circles), Indep (grey circles), and Null Model (green circles). Plot (a) shows results for the 'Potts' model, and plot (b) shows results for the 'MSA-T' model. In both plots, r_{20} decreases as Pattern Length increases. The 'Indep' model consistently shows the lowest r_{20} values, while the 'Potts' and 'MSA-T' models show higher values. The 'Null Model' is represented by a dashed line in both plots.

Pattern Length	(a) Potts r_{20}	(a) MSA-T r_{20}	(a) Indep r_{20}	(a) Null Model r_{20}	(b) Potts r_{20}	(b) MSA-T r_{20}	(b) Indep r_{20}	(b) Null Model r_{20}
2	0.98	0.85	0.95	0.98	0.98	0.85	0.95	0.98
3	0.95	0.75	0.85	0.95	0.95	0.75	0.85	0.95
4	0.85	0.65	0.65	0.85	0.85	0.65	0.65	0.85
5	0.65	0.55	0.45	0.65	0.65	0.55	0.45	0.65
6	0.50	0.45	0.35	0.50	0.50	0.45	0.35	0.50
7	0.40	0.35	0.25	0.40	0.40	0.35	0.25	0.40
8	0.35	0.30	0.20	0.35	0.35	0.30	0.20	0.35

The filtered natural MSAs have limited numbers of sequences (12.9K for RR domain), which causes significant finite-sampling error at high orders of marginal. To bypass this limitation, we conducted a “synthetic” test with large training and reference MSAs of 6M sequences. Here, we first trained an initial Potts model on filtered natural MSA and generated two synthetic 6M MSAs from it to serve as reference and training MSAs in the next step. A new Potts, MSA-T (as implemented in Ref. [23]), and an Independent model were trained on this synthetic i.i.d. training dataset. The result shown in Fig. 4(a) supports that the Potts model outperforms MSA-T when the input sequences are i.i.d. even at higher orders of marginal, which are better probed in this test. Interestingly, MSA-T performed worse than a site-independent

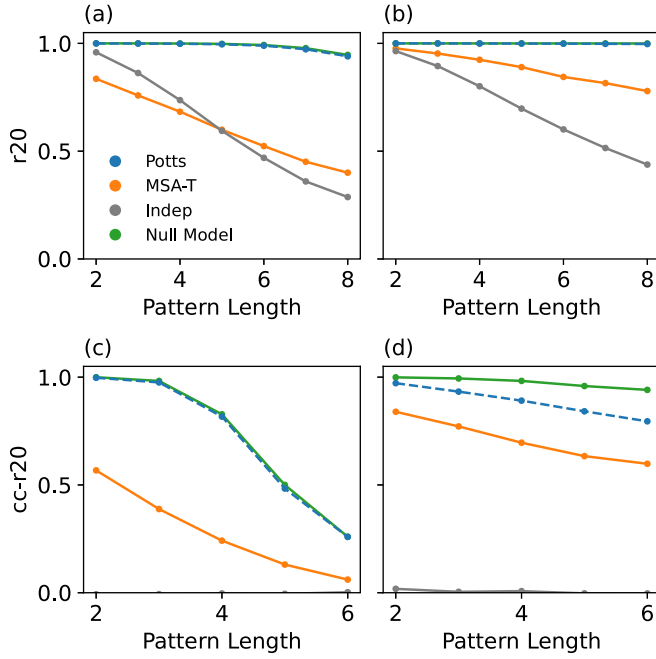


FIG. 4. “Synthetic” GPSM performance test for the RR domain in which large (6M) training and reference MSAs are produced by an initial GPSM, which is a Potts model in panels (a) and (c), and MSA-T in panels (b) and (d). The r_{20} metric is used in panels (a) and (b), and a $cc-r_{20}$ metric in panels (c) and (d).

model for low orders but slightly outperformed it at higher orders in the case of a synthetic Potts process. This synthetic test might not accurately represent how GPSMs would perform on natural datasets because the reference MSA was generated by the Potts model itself, and the refit Potts model should have zero specification error by construction. To address this, we performed another test in which the training and reference MSAs were generated by MSA-T from the natural dataset, showing in Fig. 4(b) that the Potts model still outperforms MSA-T. If MSA-T captures biophysical constraints in natural MSAs, which the Potts model cannot, we would instead expect lower Potts model performance. Interestingly, MSA-T is unable to reproduce the HOMs when trained on its own generated MSAs, as the r_{20} metric at higher order is lower than both the null expectation and the Potts result. We investigated this by testing the alternative generation algorithms presented in Ref. [23], as well as changing the acceptance rate p and other variations, but always found qualitatively similar results suggesting a general limitation of MSA-T.

An even more stringent test of the GPSMs is to measure a “connected correlation r_{20} ” ($cc-r_{20}$) metric that compares connected correlations, which are higher-order generalizations of the pairwise amino-acid covariation values $C_{\alpha\beta}^{ij} = f_{\alpha\beta}^{ij} - f_{\alpha}^i f_{\beta}^j$ [12]. A site-independent model should have zero connected correlation at all orders. In Figs. 4(c) and 4(d), MSA-T scores significantly worsen using $cc-r_{20}$ than the Potts model. Interestingly, in Fig. 4(d), where synthetic data were generated by MSA-T, the Potts model does not match the null expectation, possibly indicating that MSA-T introduces statistical patterns beyond pairwise interactions. However, in natural sequence tests (Fig. 3), the Potts model matched the null expectation,

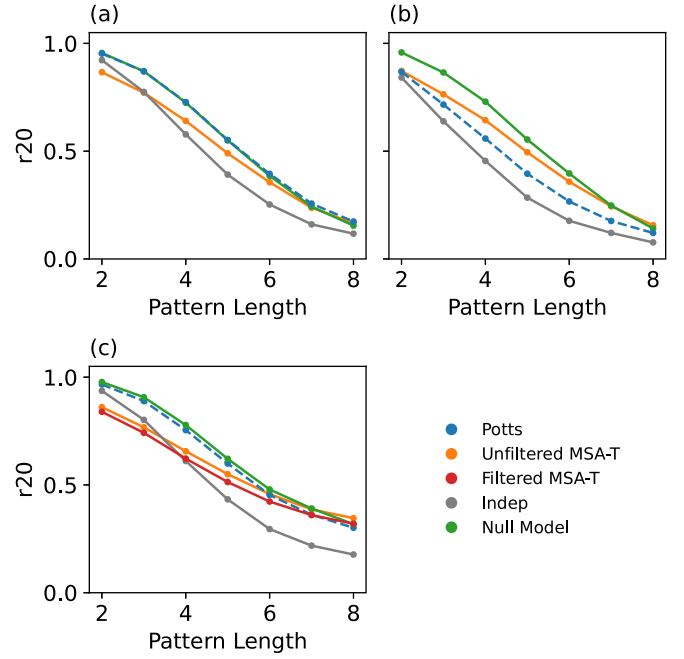


FIG. 5. Tests of the impact of identity filtering for the RR-domain family. (a) Models using 30K unfiltered natural sequences for both training and evaluation. (b) The Potts and Independent models trained on 6K filtered natural sequences; MSA-T was trained on 30K unfiltered sequences. The reference MSA is 30K unfiltered. (c) r_{20} value for all models, evaluated against 6K filtered natural sequences.

suggesting such patterns may not exist in natural datasets. MSA-T also tends to generate less variation than the Potts model, explaining why the null result is larger in Fig. 4(c) compared to Fig. 4(d).

VI. THE POTTS MODEL OUTPERFORMS MSA-T WHEN TRAINED USING PHYLOGENETICALLY REDUNDANT MSAs

In Ref. [23], it was suggested that MSA-T may have an advantage by being insensitive to phylogenetic structure due to its columnwise attention layers, avoiding the need for identity filtering that has high computational complexity $O(N^2)$ for MSA depth N . In Ref. [23], MSA-T was trained and evaluated on randomly divided MSAs without identity filtering, and the results showed that it outperformed the Potts model, according to r_{20} tested against the unfiltered reference MSA, which we reproduced in Fig. 5(b). However, the Potts model used in that test implicitly performed identity filtering as an internal step. While this can be reasonable if treating the Potts software as a black-box GPSM, it should be expected that a model trained on a filtered MSA will have a lower r_{20} when evaluated with unfiltered reference MSA. This suggests to test performance when the Potts model is both trained and evaluated on unfiltered MSAs. We find that the Potts model outperforms MSA-T in this case [Fig. 5(a)], suggesting that the Potts model is able to model the component of correlations caused by phylogenetic clusters, to some degree. We expect that such correlations are not biophysical in origin, so that

while intriguing this test is inappropriate for testing which GPSM best captures biophysical constraints.

Instead, we suggest that using filtered MSAs for model evaluation will give the best measure of a GPSM's ability to capture the underlying biophysical fitness function, as this will minimize the effects of phylogenetic clustering that introduce sequence correlations driven by nonselective parameters like speciation rates and experimental sampling bias, as discussed above. In Fig. 5(c), we compare multiple GPSMs using a filtered reference MSA, and in particular compare MSA-T performance using either filtered or unfiltered training MSAs. MSA-T shows similar performance either way, suggesting that it implicitly corrects for phylogenetic structure in the training MSA, if present. However, when the Potts model is trained using a filtered MSA, it generally outperforms MSA-T and closely matches the null expectation for a well-specified model. This supports the conclusion that the Potts model more accurately captures features of the underlying biophysical fitness function, as measured by r_{20} , than MSA-T.

In addition to the r_{20} analysis, we evaluated simpler one-site statistics for the filtered training MSA and the model-generated MSAs from it, using sequence logos generated by the WebLogo [29,30] tool, and per-site entropy for all models. The Potts and Independent models very closely match the conservation patterns of the natural MSA, while MSA-T exhibits overconservation and lower entropy [see Figs. S3(a), S3(b), and S4 in the Supplemental Material [25]]. These results further support the Potts model's stronger ability to reproduce the statistical features of the training MSA.

VII. DISCUSSION

The impact of phylogenetic relationships on GPSMs and protein covariation analysis has been recognized since early Potts studies [31–34]. Various methods have been proposed to address its confounding effects, such as the “average product correction” [35,36] and identity weighting [4], or in profile HMMs used by HMMER [37], which are a form of GPSMs, by weighting like the Henikoff scheme [38]. Spectral decomposition of the pairwise covariation matrix [39], central to Potts inference, shows that its low eigenmodes are influenced by phylogeny, while high eigenmodes capture biophysical mutational couplings due to epistasis, and that Potts inference is insensitive to low eigenmodes. This insensitivity has also been found empirically when predicting biophysical “contacts” in proteins [40]. This suggests that the Potts model may accurately model biophysical interactions even if identity filtering does not completely account for phylogenetic clustering. MSA-T is also designed to account for phylogeny through column attention heads, and it has been found that some attention heads effectively detect sequence relationships [11,21].

These results are consistent with previous studies comparing the ability of Potts models and other GPSMs to capture other aspects of protein data, including contacts in observed three-dimensional structures of proteins [41,42], experimental fitnesses, or fitness changes upon mutation [16,43], which find the architecturally simple Potts model performs favorably. For instance, a previous comparison found that the Potts model

outperforms MSA-T for contact prediction if the input data have phylogenetic structure removed [11].

Here, we showed the model's ability to reproduce higher-order sequence covariation, which is a necessary but not sufficient condition for GPSMs to produce functional, natural-like proteins. However, it has been shown that sequences generated computationally using a Potts model are functional *in vivo* [44], and similarly for other recent generative models [45–47]. Additionally, we observe that sequences generated from the Potts model, MSA-T, and the Independent model are all predicted to fold appropriately by the structure-prediction program AlphaFold [48] (see the Supplemental Material [25]). While this suggests that these sequences are foldable, experimental validation is essential to assess whether the generated sequences are fully functional.

We hypothesize that the Potts model outperforms MSA-T in capturing biophysical constraints because (1) it is trained on a specific protein family, while MSA-T is trained on all families; (2) it is directly trained to reproduce pairwise sequence statistics, whereas MSA-T is trained for a masked learning task and so its predictions of marginals are unsupervised; and (3) the Potts model generates sequences with the same diversity as the training MSA, while MSA-T has a free parameter (“replacement rate”) making unclear which value to choose [23]. The two models capture higher-order correlations in potentially different ways: The Potts model is able to model higher-order covariation through networks and chains of pairwise interactions, which may be sufficient to represent a broad class of higher-order correlations, whereas MSA-T leverages deep-learning network layers with nonlinear activations that can, in principle, capture covariation inaccessible to the Potts model. These findings imply that, despite its architectural simplicity, the Potts model best captures functional and structural protein constraints and highlights the importance of carefully decomposing the origins of covariation, not only when training GPSMs, but also during evaluation, and in their practical use for understanding the protein biophysics.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health Grant No. R35 GM132090 (R.M.L.) and by NIH Computer Equipment Grant No. OD020095. Gratitude is also expressed to the OWLSNEST high-performance cluster at Temple University for its computing support in this project.

DATA AVAILABILITY

For our study, the RR-domain and kinase MSAs were constructed from the PF00072 and PF00069 seed MSAs from the Pfam database [49], respectively. We used the Mi3-GPU software package to generate the synthetic sequences from the Potts model in our study, which is publicly accessible via its GitHub repository: [50]. To generate the synthetic sequences using MSA-T, we used the publicly available package available via GitHub at [51], which is also archived at Zenodo: [52]. Furthermore, the necessary scripts, data, and files to reproduce the results reported in our study are provided in the publicly accessible Zenodo repository: [53] and on GitHub: [54].

APPENDIX: EXTENDED METHODS

1. Natural MSA dataset preparation

For the RR-domain and protein-kinase families, we constructed MSAs using HHblits [55] to search the Uniclust database version 2023_02 [56]. For RR domain, we used the PF00072 seed MSA from Pfam database [49]. For kinase, we used the PF00069 seed MSA from Pfam [49], with pseudogenes removed as performed in Ref. [12], resulting in 73 062 raw sequences for RR domain and 291 731 raw sequences for kKinase.

2. Natural MSA preprocessing

We used identity filtering with a 50% identity threshold, resulting in MSAs of 12 906 sequences for RR domain and 20 000 sequences for kinase. The RR domain was split into 6000 sequence training and reference MSAs, and the kinase into 10 000 sequence training and reference MSAs.

When dividing unfiltered natural MSAs into training and reference sets, there can be statistical dependencies that may lead to underestimating out-of-sample error and overestimating r_{20} . To address this, we filter the MSA at a 50% identity threshold before splitting into training and reference halves. We then assign each unfiltered sequence in the original natural MSA to either training or reference set, depending on most similar sequence, ensuring closely related sequences are grouped together. This procedure minimally affects r_{20} .

3. Model training

For MSA-T, we tested both “standard” and “alternative” sequence generation methods from Ref. [23], with some modifications.

In the standard method, we divided the input MSA into 600 sequence batches and iteratively performed the masked MSA prediction task for 200 rounds per batch with a masking rate of $p = 0.1$ and default options, using the “greedy” sampling strategy. Sequences with uninitialized characters (*cls*) were discarded, and new sequences were regenerated to consist only of amino acid characters or gaps. Although convergence measures plateaued after 200 rounds, as noted in Ref. [23], continued iteration led to an accumulation of “gap” characters, especially at the sequence ends, which pronounced more at higher p values. We attribute this artifact in MSA-T

predictions to missing terminal sequences in the training data, due to shotgun sequencing, bioinformatics misannotation, or evolutionary length variation in protein termini. To address this effect biasing the higher-order correlations, we modified the iterative masking procedure so that gap characters were never masked and never generated at masked positions, preserving the input MSA’s gap characters. We computed bivariate marginals of input MSA with a pseudocount as described in Ref. [24] of scale $1/N$, where N is the input MSA depth. For training of the Potts model, we used the Mi3-GPU inference software [24].

In the “alternative” generation procedure, each sequence is iterated separately with 599 randomly drawn sequences from the input MSA, using the “logits” masked sampling strategy. We modified this to preserve gap structure and fix issues where masked positions were treated as unmasked.

4. Higher-order marginal

We use the HOM_r20 package [12] to calculate the precision of the model in reconstructing the higher-order marginals (HOMs) of MSA. We calculate the r_{20} value [12] for each HOM of the second to eighth order using 3000 randomly selected column sets of natural and evaluation MSAs. We compute the 20 most frequent amino acid subsequences for each position in training and evaluation MSAs, and then calculate the Pearson correlation (r) between these frequencies. The average r , called the r_{20} metric, indicates how well the generated MSA reconstructs the sequence statistics and higher-order mutational patterns of the natural MSA.

5. Synthetic MSA generation and analysis

We generate “synthetic” evaluation MSA datasets using models generatively. First, with the protein-kinase Potts model fit to the 12.9K filtered MSA of the RR Domain, we created 6M “synthetic training” MSAs and two sets of 6M “synthetic reference” MSAs, which were used to construct the null model, using Mi3-GPU via Markov Chain Monte Carlo (MCMC), an algorithm used to draw samples from a given probability distribution [24]. We then trained a new Potts model, as well as MSA Transformer (as implemented in Ref. [23]) and Independent models on the 6M synthetic training MSA and generated synthetic “evaluation” MSAs containing 6M sequences for all models. We repeated a similar process for the MSA Transformer using the same input as shown in Fig. 2. This approach addresses finite sampling limitations by generating MSAs with any desired number of sequences.

-
- [1] R. M. Levy, A. Haldane, and W. F. Flynn, Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness, *Curr. Opin. Struct. Biol.* **43**, 55 (2017).
 - [2] T. Bepler and B. Berger, Learning the protein language: Evolution, structure, and function, *Cell Syst.* **12**, 654 (2021).
 - [3] D. De Juan, F. Pazos, and A. Valencia, Emerging methods in protein co-evolution, *Nat. Rev. Genet.* **14**, 249 (2013).
 - [4] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing, *Proc. Natl. Acad. Sci. USA* **106**, 67 (2009).
 - [5] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, *Proc. Natl. Acad. Sci. USA* **108**, E1293 (2011).
 - [6] T. Mora and W. Bialek, Are biological systems poised at criticality? *J. Stat. Phys.* **144**, 268 (2011).
 - [7] A. Haldane, W. F. Flynn, P. He, R. Vijayan, and R. M. Levy, Structural propensities of kinase family proteins from a Potts model of residue co-variation, *Protein Sci.* **25**, 1378 (2016).

- [8] A. Haldane and R. M. Levy, Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation, *Phys. Rev. E* **99**, 032405 (2019).
- [9] J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, and M. Weigt, Efficient generative modeling of protein sequences using simple autoregressive models, *Nat. Commun.* **12**, 5800 (2021).
- [10] R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives, MSA transformer, in *International Conference on Machine Learning*, Proceedings of Machine Learning Research (PMLR, 2021), pp. 8844–8856.
- [11] U. Lupo, D. Sgarbossa, and A. F. Bitbol, Protein language models trained on multiple sequence alignments learn phylogenetic relationships, *Nat. Commun.* **13**, 6298 (2022).
- [12] F. McGee, S. Hauri, Q. Novinger, S. Vucetic, R. M. Levy, V. Carnevale, and A. Haldane, The generative capacity of probabilistic protein sequence models, *Nat. Commun.* **12**, 6302 (2021).
- [13] J. Martin, M. Lequerica Mateos, J. N. Onuchic, I. Coluzza, and F. Morcos, Machine learning in biological physics: From biomolecular prediction to design, *Proc. Natl. Acad. Sci. USA* **121**, e2311807121 (2024).
- [14] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett 30 (Curran Associates, Inc., Long Beach, California, USA, 2017).
- [16] A. Hawkins-Hooker, D. T. Jones, and B. Paige, MSA-conditioned generative protein language models for fitness landscape modelling and design, in *Machine Learning for Structural Biology Workshop, NeurIPS* (2021).
- [17] Y. Hong, J. Song, J. Ko, J. Lee, and W.-H. Shin, S-Pred: Protein structural property prediction using MSA Transformer, *Sci. Rep.* **12**, 13891 (2022).
- [18] C. Ma, H. Zhao, L. Zheng, J. Xin, Q. Li, L. Wu, Z. Deng, Y. Lu, Q. Liu, and L. Kong, Retrieved sequence augmentation for protein representation learning, *arXiv:2302.12563*.
- [19] B. Almalki and L. Liao, Tmhc-msat: Accurate prediction of inter-helical residue contacts in transmembrane proteins using MSA Transformer, in *Proceedings of the 16th International Conference on Bioinformatics and Computational Biology (BIOB-2024)*, edited by H. Al-Mubaid, T. Aldwairi, and O. Eulenstein, EPiC Series in Computing (EasyChair, New Orleans, Louisiana, USA, 2024), Vol. 101, pp. 1–10.
- [20] F. Cuturello, M. Celoria, A. Ansuini, and A. Cazzaniga, Enhancing predictions of protein stability changes induced by single mutations using MSA-based language models, *Bioinformatics* **40**, btae447 (2024).
- [21] R. Chen, G. Foley, and M. Boden, Learning the language of phylogeny with MSA transformer, *bioRxiv* (2024), doi: <https://doi.org/10.1101/2024.12.18.629037>.
- [22] Y. Chen, G. Chen, and C. Y.-C. Chen, MFTrans: A multi-feature transformer network for protein secondary structure prediction, *Int. J. Biol. Macromol.* **267**, 131311 (2024).
- [23] D. Sgarbossa, U. Lupo, and A. F. Bitbol, Generative power of a protein language model trained on multiple sequence alignments, *eLife* **12**, e79854 (2023).
- [24] A. Haldane and R. M. Levy, Mi3-GPU: Mcmc-based inverse ising inference on gpus for protein covariation analysis, *Comput. Phys. Commun.* **260**, 107312 (2021).
- [25] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/g5cx-1vhj> for additional figures and analyses supporting the main text, which also includes Refs. [52,53].
- [26] B. Rost, Protein structures sustain evolutionary drift, *Folding Des.* **2**, S19 (1997).
- [27] B. Rost, Twilight zone of protein sequence alignments, *Protein Engineering* **12**, 85 (1999).
- [28] R. D. Finn, A. Bateman, J. Clements, P. Coghill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry *et al.*, Pfam: The protein families database, *Nucleic Acids Res.* **42**, D222 (2014).
- [29] T. D. Schneider and R. M. Stephens, Sequence logos: A new way to display consensus sequences, *Nucleic Acids Res.* **18**, 6097 (1990).
- [30] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, WebLogo: A sequence logo generator, *Genome Res.* **14**, 1188 (2004).
- [31] A. S. Lapedes, B. G. Giraud, L. Liu, and G. D. Stormo, Correlated mutations in models of protein sequences: Phylogenetic and structural effects, *Stat. Mol. Biol. Genet.* **33**, 236 (1999).
- [32] J. Y. Dutheil, Detecting coevolving positions in a molecule: Why and how to account for phylogeny, *Brief. Bioinform.* **13**, 228 (2012).
- [33] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, Transformer protein language models are unsupervised structure learners, in *International Conference on Learning Representations* (2021), doi: <https://doi.org/10.1101/2020.12.15.422761>.
- [34] A. Colavin, E. Atolia, A. F. Bitbol, and K. C. Huang, Extracting phylogenetic dimensions of coevolution reveals hidden functional signals, *Sci. Rep.* **12**, 820 (2022).
- [35] 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**, 333 (2008).
- [36] C. M. Buslje, J. Santos, J. M. Delfino, and M. Nielsen, Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information, *Bioinformatics* **25**, 1125 (2009).
- [37] R. D. Finn, J. Clements, and S. R. Eddy, HMMER web server: Interactive sequence similarity searching, *Nucleic Acids Res.* **39**, W29 (2011).
- [38] S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks., *Proc. Natl. Acad. Sci. USA* **89**, 10915 (1992).
- [39] C. Qin and L. J. Colwell, Power law tails in phylogenetic systems, *Proc. Natl. Acad. Sci. USA* **115**, 690 (2018).
- [40] E. R. Hortal and M. Weigt, On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins, *PLoS Comput. Biol.* **17**, 1 (2021).
- [41] N. Bhattacharya, N. Thomas, R. Rao, J. Dauparas, P. K. Koo, D. Baker, Y. S. Song, and S. Ovchinnikov, Interpreting Potts and transformer protein models through the lens of simplified attention, in *Pacific Symposium on Biocomputing 2022* (World Scientific, 2021), pp. 34–45.
- [42] Y. Hong, J. Lee, and J. Ko, A-Prot: Protein structure modeling using MSA transformer, *BMC Bioinf.* **23**, 93 (2022).

- [43] S. Cocco, L. Posani, and R. Monasson, Functional effects of mutations in proteins can be predicted and interpreted by guided selection of sequence covariation information, *Proc. Natl. Acad. Sci. USA* **121**, e2312335121 (2024).
- [44] S. Alvarez, C. M. Nartey, N. Mercado, J. A. de la Paz, T. Huseinbegovic, and F. Morcos, *In vivo* functional phenotypes from a computational epistatic model of evolution, *Proc. Natl. Acad. Sci. USA* **121**, e2308895121 (2024).
- [45] B. Zhou, L. Zheng, B. Wu, K. Yi, B. Zhong, Y. Tan, Q. Liu, P. Liò, and L. Hong, A conditional protein diffusion model generates artificial programmable endonuclease sequences with enhanced activity, *Cell Discov.* **10**, 95 (2024).
- [46] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos Jr, C. Xiong, Z. Z. Sun, R. Socher *et al.*, Large language models generate functional protein sequences across diverse families, *Nat. Biotechnol.* **41**, 1099 (2023).
- [47] Z. Wu, K. E. Johnston, F. H. Arnold, and K. K. Yang, Protein sequence design with deep generative models, *Curr. Opin. Chem. Biol.* **65**, 18 (2021).
- [48] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, ColabFold: Making protein folding accessible to all, *Nat. Methods* **19**, 679 (2022).
- [49] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart *et al.*, The Pfam protein families database in 2019, *Nucleic Acids Res.* **47**, D427 (2019).
- [50] <https://github.com/ahaldane/Mi3-GPU>.
- [51] https://github.com/Bitbol-Lab/Iterative_masking.
- [52] <https://doi.org/10.5281/zenodo.7684052>.
- [53] K. Khatri, R. M. Levy, and A. Haldane, Phylogenetic corrections and higher-order sequence statistics in protein families: The Potts vs multiple sequence alignment transformer machine learning models, Zenodo (2025), <https://doi.org/10.5281/zenodo.15876209>
- [54] K. Khatri, Potts vs MSA Transformer r_20 analysis, 2025, https://github.com/kisan-khatri/Potts_vs_MSA_Transformer-r20_Analysis.
- [55] M. Remmert, A. Biegert, A. Hauser, and J. Söding, HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment, *Nat. Methods* **9**, 173 (2012).
- [56] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger, Uniclust databases of clustered and deeply annotated protein sequences and alignments, *Nucleic Acids Res.* **45**, D170 (2017).