

# Phylogenetic Corrections and Higher-Order Sequence Statistics in Protein Families: The Potts Model vs MSA Transformer

Kisan Khatri,<sup>1</sup> Ronald M. Levy,<sup>2</sup> and Allan Haldane<sup>3,\*</sup>

<sup>1</sup>*Department of Physics and Center for Biophysics and Computational Biology, Temple University, Philadelphia, PA 19122, USA*

<sup>2</sup>*Department of Chemistry and Center for Biophysics and Computational Biology, Temple University, Philadelphia, PA 19122, USA*

<sup>3</sup>*Department of Physics and Center for Biophysics and Computational Biology, Temple University, Philadelphia, PA 19122, USA*

Recent generative 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 which 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 biological sequence constraints. We find that the model performance depends greatly on the treatment of phylogenetic relationships between the sequences, which can induce non-biophysical mutational covariation in MSAs. When using explicit corrections for phylogenetic dependencies, we find the Potts model outperforms MSA-T in detecting epistatic interactions of biophysical origin.

*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 (GPSM) 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 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, 7]. This raises the question of the GPSM functional form that best describes protein biophysics and how to measure this. Here we examine two leading GPSMs in recent focus, the Potts model[2-8] and the MSA Transformer (MSA-T)[9, 10]. We suggest that certain statistical characteristics of individual protein families provide the most reliable metrics to measure GPSM performance if they exclude the biasing effects of phylogenetic relationships between sequences, which do not originate from the fundamental biophysical properties of the protein family. We find that the Potts model better captures such a statistic, the higher-order MSA statistics which arise due to the potentially large epistatic networks within proteins, when the impact of phylogenetic relationships is carefully accounted for.

*Effect of Phylogenetic relationships on GPSM evaluation* - Mutational covariation in protein sequences provides a highly informative statistical signal that accurate GPSMs must capture. These arise from multiple factors including: 1) High-fitness sequence motifs and epistasis (mutational cooperativity) underlying biophysical func-

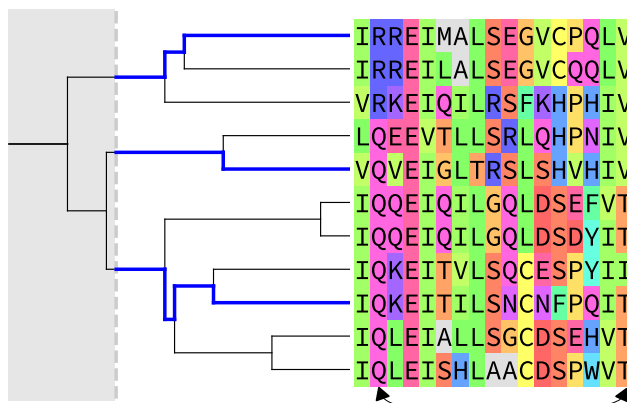


Figure 1. Phylogenetic relationships between sequences in an MSA result in a spurious mutational correlation due to common ancestry, for R/V and Q/T combinations at the illustrated column-pair. Sequence pairs greater than an identity cut-off (diverging to the left of the gray dotted line) approximate *i.i.d.* samples, so that identity filtering to retain 3 sequences labeled in blue gives an unbiased sample showing no correlation with equal frequency of R/V, Q/V and Q/T combinations.

tion 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, and 3) 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.

Despite their quite different architectures and complexity, the Potts model and MSA-T exhibit fundamental commonalities in their design to account for these patterns[11]. The Potts model is an interpretable, physics-inspired machine learning model fit to a single protein family, inspired by spin-glasses. A Potts model for 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  $\alpha, \beta$  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 [12] masked language neural network attention model [13] trained on MSAs of all available protein families[14] with about  $10^4$  times more parameters, recently used for protein structure and evolution analyses[15–21]. The Potts model has been found to have superior generative accuracy to some other GPSMs including Variational Auto-Encoders and site-independent models [22], 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 unclear.

GPSMs must distinguish the causes of covariation[24]. 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 column-wise attention layer[10, 20], 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 this in a generative fashion for a single protein family by repeatedly masking and resampling an input MSA of that family to produce

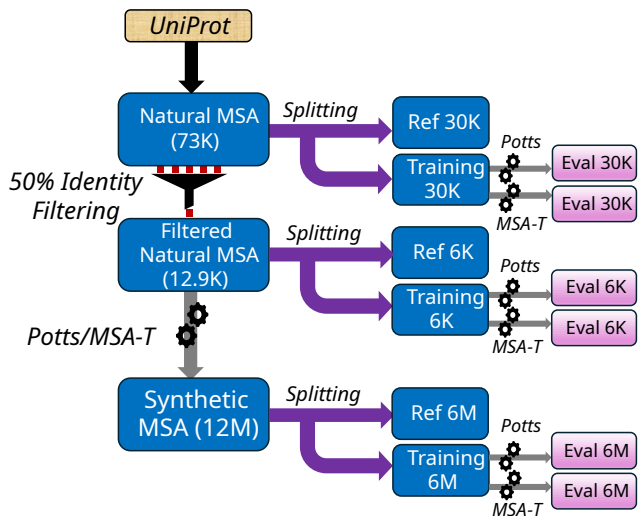


Figure 2. Overview of statistical tests carried out using the Potts model and MSA-T, to isolate different forms of statistical error. Boxes represent MSAs with different amounts of sequences shown for the RR-domain family, which are then filtered, split, or used to train and generate from the GPSMs (arrows). Our tests measure the statistical difference between the “evaluation” MSAs and the corresponding “reference” MSAs.

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 pre-trained parameters  $\theta$ , an input MSA  $M_F$  for a protein family  $F$ , and a masking frequency parameter  $p$ .

One goal of this study is to control for how phylogenetic clusters affects 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.

*Statistical methodology* - We perform tests of GPSM performance with and without accounting for phylogeny as outlined in Figure 2. We generate sequences from MSA-T using the default method of Ref. [23] using pre-computed parameters  $\theta$ , and train the Potts model using the high-accuracy Mi3-GPU method[25]. We also test a very simple GPSM, the Independent model, with distribution  $P(S) = \prod_i f_{s_i}^i$  where  $f_{\alpha}^i$  are single-site frequencies for amino-acid  $\alpha$  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. We train the GPSMs on various MSA data described in subsequent sections, generate new synthetic 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[22]. 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[22], 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 (PF000720), which was previously studied to test MSA-T predictions of higher-order sequence statistics[23], and protein-kinase (PF00069)[22]. The details of the methodology are provided in the “End Matter”.

The  $r_{20}$  metric measures how well the GPSM predicts the frequency of non-contiguous amino acid “words” of length  $n$  as described in detail previously[22]. In summary, for each order  $n$ , 3000 randomly chosen sets of positions of size  $n$  are evaluated. For each set of positions, the top 20 most common “words” at these positions in the reference MSA are found, and the frequency of these words are also computed in both the reference MSA and in the GPSM-generated evaluation MSA. The  $r_{20}$  metric is the Pearson correlation between the 20 reference and 20 evaluation frequencies, averaged over the 3000 sets. Focusing on only the top 20 most frequent words restricts the computation to statistically reliable values. This metric measures the ability of the GPSM to model complex chains and networks of interactions within proteins.

To perform identify filtering of MSAs we iteratively find pairs of sequences in the MSA are more than 50% similar and randomly drop one of the two, because in most protein families distantly related sequences have 10%-50% identity, while  $> 50\%$  suggests recent ancestry and non-*i.i.d.* sequences. This largely eliminates the influence of phylogenetic clusters on MSA statistics so that most covariation in the training MSA is biophysically induced. For RR-domain, we have 73K sequences, which become filtered to 12.9K sequences, then split into 6K reference and training. For protein-kinase, we have 292K sequence which become filtered to 20K sequences. We also performed our analyses using MSAs filtered at 60% and 90% cutoffs to test if the results were sensitive to cutoff, finding qualitatively they were not.

*The Potts model outperforms MSA-T after Phylogenetic Pre-processing-* We first tested the performance of the model after filtering the training and reference MSAs, corresponding to the middle section of Figure 2. In Figures 3(a) (RR domain) and 3(b) (kinase), we show that the Potts model outperforms MSA-T in this test.

The  $r_{20}$  metric is lower at higher-orders because of greater finite sampling error when measuring the smaller frequencies at these orders, and not because of reduced model accuracy[22]. The maximum attainable  $r_{20}$  metric for a well-specified model subject only to finite-sampling

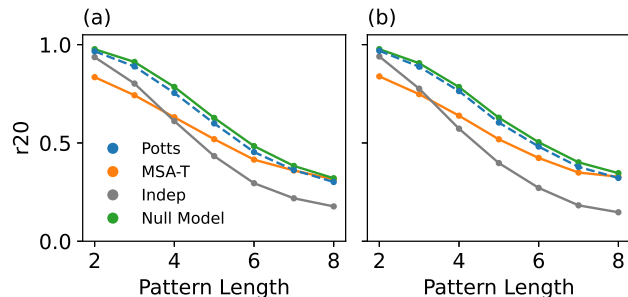


Figure 3. “Natural” GPSM performance test in which the training and reference MSAs are natural sequences filtered by sequence identity to eliminate phylogenetic redundancy, and evaluated using the  $r_{20}$  metric for (a) RR-domain (MSAs of 6K) (b) Kinase Protein (MSAs of 10K).

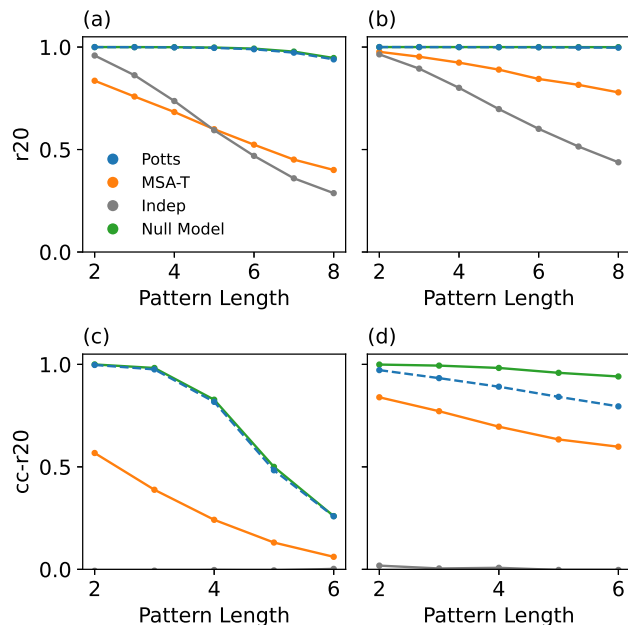


Figure 4. “Synthetic” GPSM performance test for the RR Domain in which large (6M) training and reference MSAs are produced by an initial GPSM, which is a Potts model in (a),(c), and MSA-T in (b),(d). The  $r_{20}$  metric is used in (a),(b), and a  $cc-r_{20}$  metric in (c) and (d).

limitations in model evaluation can be found by measuring  $r_{20}$  between the training MSA and reference MSA. The Potts model results closely match this null model value, unlike MSA-T.

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, and Independent model were trained on this synthetic *i.i.d.* training dataset. The result shown in Figure 4a 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 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 re-fit 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 Figure 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 are 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 which compares connected correlations, which are higher-order generalizations of the pairwise amino-acid covariation values  $C_{\alpha\beta}^{ij} = f_{\alpha\beta}^{ij} - f_{\alpha}^i f_{\beta}^j$  [22]. A site-independent model should have zero connected correlation at all orders. In Figures 4(c) and 4(d), MSA-T scores significantly worse using  $cc-r_{20}$  than the Potts model. Interestingly, in 4(d), where synthetic data was generated by MSA-T, the Potts model does not match the null expectation, possibly indicating MSA-T introduces statistical patterns beyond pairwise interactions. However, in natural sequence tests (Fig 3), the Potts model matched the null expectation, 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 Figure 4(c) compared to 4(d).

*The Potts model outperforms MSA-T when trained using phylogenetically redundant MSAs* - In [23], it was suggested that MSA-T may have an advantage by being insensitive to phylogenetic structure due to its column-wise attention layers, avoiding the need for identity-filtering which has high computational complexity  $O(N^2)$  for MSA depth  $N$ . In [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 un-

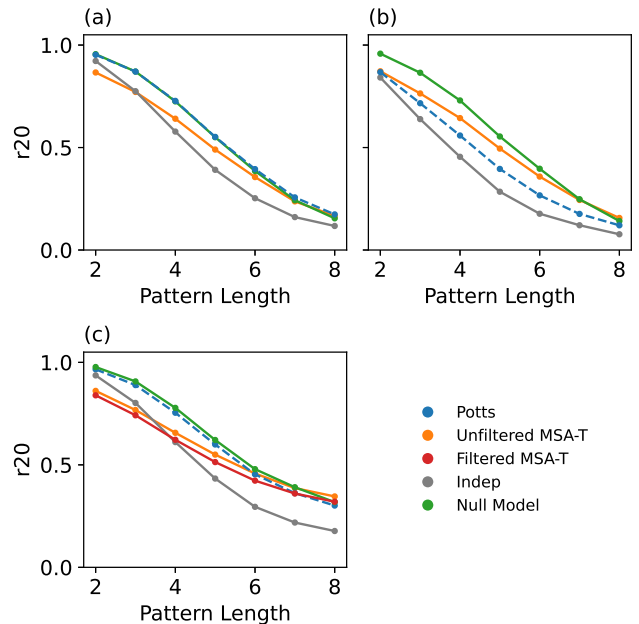


Figure 5. Tests of the impact of identity filtering, for the RR-domain family (a) Models using 30K unfiltered natural sequences for both training and evaluation. (b) The Potts and Independent models trained on 6K filtered natural sequences; MSA-T was trained on 30K unfiltered sequences. The reference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

filtered reference MSA, which we reproduced in Figure 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 the Potts model outperforms MSA-T in this case (Figure 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 a GPSM’s ability to capture the underlying biophysical fitness function, as this will minimize the effects of phylogenetic clustering which introduce sequence correlations driven by non-selective parameters like speciation rates and experimental sampling bias, as discussed above. In Figure 5(c) we compare multiple GPSMs using a filtered reference MSA, and in particular compare MSA-T performance using either filtered and unfiltered training MSAs. MSA-T

shows similar performance either way, suggesting 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.

*Discussion* - The impact of phylogenetic relationships on GPSMs and protein covariation analysis has been recognized since early Potts studies[26–29]. Various methods have been proposed to address its confounding effects, such as the “Average Product Correction” (APC)[30, 31] and identity-weighting [2], or, in profile-HMMs used by HMMER[32] which are a form of GPSMs, by weighting like the Henikoff scheme[33]. Spectral decomposition of the pairwise covariation matrix [34], central to Potts inference, shows 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 [35]. This suggests 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[10, 20].

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 3d structures of proteins[36, 37], experimental fitnesses or fitness changes upon mutation[15, 38] 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 has phylogenetic structure removed[10].

We hypothesize 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]. These findings imply the Potts model best captures functional and structural protein constraints despite its architectural simplicity, and highlight the importance of carefully decomposing the origins of covariation, not only when training GPSMs

but also during evaluation and in their practical use in understanding the biophysical properties of proteins.

*Acknowledgements* - This research was supported by National Institutes of Health (NIH) grant number R35-GM132090, and by NIH Computer Equipment Grant (OD020095). Gratitude is also expressed to the OWL-SNEST high performance cluster at Temple University for its computing support in this project.

---

\* [allan.haldane@temple.edu](mailto:allan.haldane@temple.edu)

- [1] T. Bepler and B. Berger, Learning the protein language: Evolution, structure, and function, *Cell Systems* **12**, 654 (2021).
- [2] 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, *Proceedings of the National Academy of Sciences* **106**, 67 (2009).
- [3] 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, *Proceedings of the National Academy of Sciences* **108**, E1293 (2011).
- [4] T. Mora and W. Bialek, Are biological systems poised at criticality?, *Journal of Statistical Physics* **144**, 268 (2011).
- [5] 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 Science* **25**, 1378 (2016).
- [6] A. Haldane and R. M. Levy, Influence of multiple-sequence-alignment depth on potts statistical models of protein covariation, *Physical Review E* **99**, 032405 (2019).
- [7] R. M. Levy, A. Haldane, and W. F. Flynn, Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness, *Current opinion in structural biology* **43**, 55 (2017).
- [8] J. Trinquier, G. Uguzzoni, A. Pagnani, F. Zamponi, and M. Weigt, Efficient generative modeling of protein sequences using simple autoregressive models, *Nature Communications* **12**, [10.1038/s41467-021-25756-4](https://doi.org/10.1038/s41467-021-25756-4) (2021).
- [9] 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* (PMLR, 2021) pp. 8844–8856.
- [10] U. Lupo, D. Sgarbossa, and A. F. Bitbol, Protein language models trained on multiple sequence alignments learn phylogenetic relationships, *Nature Communications* **13**, [10.1038/s41467-022-34032-y](https://doi.org/10.1038/s41467-022-34032-y) (2022).
- [11] J. Martin, M. Lequerica Mateos, J. N. Onuchic, I. Coluzza, and F. Morcos, Machine learning in biological physics: From biomolecular prediction to design, *Proceedings of the National Academy of Sciences* **121**, e2311807121 (2024).
- [12] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *nature* **521**, 436 (2015).

- [13] 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* **30** (2017).
- [14] R. Rao, J. Liu, R. Verkuil, J. Meier, J. F. Canny, P. Abbeel, T. Sercu, and A. Rives, *Msa transformer* (2021).
- [15] 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).
- [16] Y. Hong, J. Song, J. Ko, J. Lee, and W.-H. Shin, S-pred: protein structural property prediction using msa transformer, *Scientific reports* **12**, 13891 (2022).
- [17] 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, *bioRxiv*, 2023 (2023).
- [18] 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*, Vol. 101 (2024) pp. 1–10.
- [19] 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).
- [20] R. Chen, G. Foley, and M. Boden, Learning the language of phylogeny with msa transformer, *bioRxiv*, 2024 (2024).
- [21] Y. Chen, G. Chen, and C. Y.-C. Chen, Mftrans: A multi-feature transformer network for protein secondary structure prediction, *International Journal of Biological Macromolecules* **267**, 131311 (2024).
- [22] F. McGee, S. Hauri, Q. Novinger, S. Vucetic, R. M. Levy, V. Carnevale, and A. Haldane, The generative capacity of probabilistic protein sequence models, *Nature Communications* **12**, 10.1038/s41467-021-26529-9 (2021).
- [23] D. Sgarbossa, U. Lupo, and A. F. Bitbol, Generative power of a protein language model trained on multiple sequence alignments, *eLife* **12**, 10.7554/eLife.79854 (2023).
- [24] D. De Juan, F. Pazos, and A. Valencia, Emerging methods in protein co-evolution, *Nature Reviews Genetics* **14**, 249 (2013).
- [25] A. Haldane and R. M. Levy, Mi3-gpu: Mcmc-based inverse ising inference on gpus for protein covariation analysis, *Computer Physics Communications* **260**, 107312 (2021).
- [26] A. S. Lapedes, B. G. Giraud, L. Liu, and G. D. Stormo, Correlated mutations in models of protein sequences: phylogenetic and structural effects, *Lecture Notes-Monograph Series*, 236 (1999).
- [27] J. Y. Dutheil, Detecting coevolving positions in a molecule: Why and how to account for phylogeny, *Briefings in Bioinformatics* **13**, 228 (2012).
- [28] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, Transformer protein language models are unsupervised structure learners, *Biorxiv*, 2020 (2020).
- [29] A. Colavin, E. Atolia, A. F. Bitbol, and K. C. Huang, Extracting phylogenetic dimensions of coevolution reveals hidden functional signals, *Scientific Reports* **12**, 10.1038/s41598-021-04260-1 (2022).
- [30] 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).
- [31] 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 co-evolving amino acid pairs using mutual information, *Bioinformatics* **25**, 1125 (2009).
- [32] R. D. Finn, J. Clements, and S. R. Eddy, HMMER web server: interactive sequence similarity searching, *Nucleic Acids Research* **39**, W29 (2011), [https://academic.oup.com/nar/article-pdf/39/suppl\\_2/W29/7628106](https://academic.oup.com/nar/article-pdf/39/suppl_2/W29/7628106)
- [33] S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks., *Proceedings of the National Academy of Sciences* **89**, 10915 (1992).
- [34] C. Qin and L. J. Colwell, Power law tails in phylogenetic systems, *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12222 (2002).
- [35] E. R. Hortal and M. Weigt, On the effect of phylogenetic correlations in coevolution-based contact prediction in proteins, *PLoS Computational Biology* **17**, 10.1371/journal.pcbi.1008957 (2021).
- [36] 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 physics, *bioRxiv*, 2021 (2021).
- [37] Y. Hong, J. Lee, and J. Ko, A-prot: protein structure modeling using msa transformer, *BMC bioinformatics* **23**, 93 (2022).
- [38] 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, *Proceedings of the National Academy of Sciences* **121**, e2312335121 (2024).
- [39] M. Remmert, A. Biegert, A. Hauser, and J. Söding, Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment, *Nature methods* **9**, 173 (2012).
- [40] 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 research* **47**, D427 (2019).
- [41] 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 Research* **45**, D170 (2016), <https://academic.oup.com/nar/article-pdf/45/D1/D170/8846789/gk>

## END MATTER: EXTENDED METHODS

### Natural MSA dataset preparation

For the RR domain and protein-kinase families, we constructed MSAs using HHblits[39] to search the Uni-clust database version 2023\_02 [41]. For RR Domain, we used the PF00072 seed MSA from Pfam database [40]; for Kinase, we used PF00069 with pseudogenes removed as performed in[22], resulting in 73,062 raw sequences for RR Domain and 291,731 raw sequences for Kinase.

### 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 6,000 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  $r_{20}$ .

### Model Training

For MSA-T, We tested both “standard” and “alternative” sequence generation methods from [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 [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 train-

ing data, due to shotgun sequencing, bioinformatics mis-annotation, 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. [25] of scale  $1/N$ , where  $N$  is the input MSA depth. For training of the Potts model, we used the Mi3-GPU inference software[25].

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.

### Higher-Order Marginal (HOM)

We use the HOM\_r20 package[22] to calculate the precision of the model in reconstructing the higher-order marginals (HOMs) of MSA. We calculate the  $r_{20}$  value [22] 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, 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.

### 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 using Mi3GPU using MCMC[25]. We then trained new Potts, MSA Transformer, 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. This approach addresses finite sampling limitations by generating MSAs with any desired number of sequences.