Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

# Interpreting Potts and Transformer Protein Models Through the Lens of Simplified Attention

Nicholas Bhattacharya[1,*], Neil Thomas[1,*],
Roshan Rao[1], Justas Dauparas[2], Peter K. Koo[3],
David Baker [2], Yun S. Song[1,4], and Sergey Ovchinnikov[5]

[1] University of California, Berkeley
[2] University of Washington
[3] Cold Spring Harbor Laboratory
[4] Chan Zuckerberg Biohub
[5] Harvard University

*E-mail: nick_bhat@berkeley.edu, nthomas@berkeley.edu*

The established approach to unsupervised protein contact prediction estimates coevolving positions using undirected graphical models. This approach trains a Potts model on a Multiple Sequence Alignment. Increasingly large Transformers are being pretrained on unlabeled, unaligned protein sequence databases and showing competitive performance on protein contact prediction. We argue that attention is a principled model of protein interactions, grounded in real properties of protein family data. We introduce an energy-based attention layer, *factored attention*, which, in a certain limit, recovers a Potts model, and use it to contrast Potts and Transformers. We show that the Transformer leverages hierarchical signal in protein family databases not captured by single-layer models. This raises the exciting possibility for the development of powerful structured models of protein family databases.

*Keywords*: Contact Prediction, Representation Learning, Language Modeling, Attention, Transformer, BERT, Markov Random Fields, Potts Models, Self-supervised learning

**Supplementary information:** Supplementary methods, figures, tables and code can be found at `https://github.com/songlab-cal/factored-attention`.

## 1. Introduction

Inferring protein structure from sequence is a longstanding problem in computational biochemistry. Potts models, a particular kind of Markov Random Field (MRF), are the predominant unsupervised method for modeling interactions between amino acids. Potts models are trained to maximize pseudolikelihood on alignments of evolutionarily related proteins.[1–3] Features derived from Potts models were the main drivers of improved performance at the CASP11 competition.[4] Potts models were subsequently used as input features for top performing supervised neural network models in CASP13.[5–7]

*Equal Contribution.

© 2021 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Inspired by the success of BERT,[8] GPT[9] and related unsupervised models in NLP, a line of work has emerged that learns features of proteins through self-supervised pretraining.[10–14] This new approach trains Transformer[15] models on large datasets of protein sequences. Pretrained model performance raises questions about the importance of data and model scale,[11,16] whether neural features compete with evolutionary features extracted by established bioinformatic methods,[12] and the benefits of transfer learning.[17–19]

In CASP14, Alphafold2 achieved breakthrough performance by replacing the Potts model with an attention-based model that directly used the MSA as input.[20] This approach was adapted subsequently in RoseTTAFold.[21] The performance of these methods established attention as state-of-the-art for extracting features from MSAs. This raises a natural question of how Potts models and attention mechanisms are related.

In this paper, we investigate the ways in which attention-based models and Potts models trained on alignments can learn meaningful interactions in biological sequence data. To do so, we introduce a simplified energy-based attention model trained on alignments, *factored attention*, which interpolates between the standard attention mechanism and Potts models. We show that factored attention can successfully share parameters across positions within a family or share amino acid features across hundreds of families.

## 2. Background

Proteins are polymers composed of amino acids and are commonly represented as strings. Along with this 1D sequence representation, each protein folds into a 3D physical structure. Physical distance between positions in 3D is often a much better indicator of functional interaction than proximity in sequence. One representation of physical distance is a *contact map* $C$, a symmetric matrix in which entry $C_{ij} = 1$ if the beta carbons[a] of $i$ and $j$ are within 8Å of one another, and 0 otherwise.

**Multiple Sequence Alignments.** To understand structure and function of a protein sequence, one typically assembles a set of its evolutionary relatives and looks for patterns within the set. A set of related sequences is referred to as a *protein family*, commonly represented by a Multiple Sequence Alignment (MSA). Gaps in aligned sequences correspond to insertions from an alignment algorithm,[22,23] ensuring that positions with similar structure and function line up for all members of the family. After aligning, sequence position carries significant evolutionary, structural, and functional information.

**Coevolutionary Analysis of Protein Families.** The observation that statistical patterns in MSAs can be used to predict couplings has been widely used to infer structure and function from protein families.[24–27]

## 3. Methods

To explore how attention and Potts models learn interactions in protein sequence data, we compare a number of unsupervised methods which learn contacts with sequence-modeling objectives. Many of these methods are based on the formalism of Markov Random Fields

[a]In the case of glycine, the alpha carbon is used.

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

(MRFs). We do not extend our analysis to *supervised* contact prediction models which take MRF features as input, as these are outside the scope of this work.

Throughout this section, $x = (x_1, \ldots, x_L)$ is a sequence of length $L$ from an alphabet of size $A$. This sequence is part of an MSA of length $L$ with $N$ total sequences. Recall that a fully-connected Pairwise MRF over $p$ variables $X_1, \ldots, X_p$ specifies a distribution

$$p_\theta(x_1, \ldots, x_p) = \frac{1}{Z} \exp\left( \sum_{i<j} E_\theta(x_i, x_j) \right), \tag{1}$$

where $Z$ is the partition function and $E_\theta(x_i, x_j)$ is an arbitrary function of $i$, $j$, $x_i$ and $x_j$. For all models below, we can introduce an explicit functional $E_\theta(x_i)$ to capture the marginal distribution of $X_i$. When introduced, we parametrize the marginal with $E_\theta(x_i) = b_{i,x_i}$ for $b \in \mathbb{R}^{L \times A}$.

### 3.1. *Potts Models*

A Potts model is a fully-connected pairwise MRF with $L$ variables, each representing a position in the MSA. An edge $(i, j)$ is parametrized with a matrix $W^{ij} \in \mathbb{R}^{A \times A}$. These matrices are organized into an order-4 tensor which form the parameters of a Potts model. Note that $W^{ij} = W^{ji}$. The energy functional of a Potts model is given through lookups, namely

$$E_\theta(x_i, x_j) = W^{ij}(x_i, x_j). \tag{2}$$

### 3.2. *Factored Attention*

Factored attention has two advantages over Potts for modeling protein families: it shares a pool of amino acid feature matrices across all positions and it estimates $\mathcal{O}(L)$ parameters instead of $\mathcal{O}(L^2)$.

**Sharing amino acid features.** Many contacts in a protein are driven by similar interactions between amino acids, such as many types of weakly polar interactions.[28,29] If two pairs of positions $(i, j)$ and $(l, m)$ are both in contact due to the same interaction, a Potts model must estimate completely separate amino acid features $W^{ij}$ and $W^{lm}$. In order to share amino acid features, we want to compute all energies from one pool of $A \times A$ feature matrices. The simplest way to accomplish this is by associating an $L \times L$ matrix $\mathcal{A}$ to every $A \times A$ feature matrix $W_V$. For $H$ such pairs $(\mathcal{A}, W_V)$, we could introduce a factorized MRF:

$$E_\theta(x_i, x_j) = \sum_{h=1}^{H} \operatorname{symm}\left( \operatorname{softmax}\left( \mathcal{A}^h \right) \right)_{ij} W_V^h(x_i, x_j). \tag{3}$$

A row-wise softmax is taken to encourage sparse interactions and aid in normalization. This model allows the pairs $(i, j)$ and $(l, m)$ to reuse a single feature $W_V^h$, assuming $\mathcal{A}_{ij}^h$ and $\mathcal{A}_{lm}^h$ are both large.

**Scaling linearly in length.** Both Potts and the factorized model in Equation 3 have $\mathcal{O}(L^2)$ parameters. However, contacts are observed to grow linearly over the wide range of protein structures currently available.[30,31] Given that the number of interactions we wish to estimate grows linearly in length, the quadratic scaling of these models can be greatly

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

improved. One way to fix this is by introducing the factorization $\mathcal{A} = W_Q W_K^T$, where $W_Q, W_K \in \mathbb{R}^{L \times d}$. We use the subscripts $Q$, $K$, and $V$ in analogy with the "Query", "Key", and "Value" nomenclature from the attention literature.[15] As before, we employ a row-wise softmax for sparsity and normalization. Combining feature sharing with linear length scaling leads to *factored attention*, defined in Equation 4.

Like Potts, factored attention is a fully-connected pairwise MRF with $L$ variables. The parameters of this model consist of $H$ triples $(W_Q, W_K, W_V)$, where $W_Q, W_K \in \mathbb{R}^{L \times d}$; $W_V \in \mathbb{R}^{A \times A}$; and $d$ is a hyperparameter. Each such triple is called a *head* and $d$ is the *head size*. Unlike a Potts model, the parameters for each edge $(i, j)$ are tied through the use of heads. The energy functional is

$$E_\theta(x_i, x_j) = \sum_{h=1}^{H} \operatorname{symm}\left(\operatorname{softmax}\left(W_Q^h W_K^{h\,T}\right)\right)_{ij} W_V^h(x_i, x_j), \tag{4}$$

where $\operatorname{symm}(M) = (M + M^T)/2$ ensures the positional interactions are symmetric.

Adding sequence-dependent interactions leads to standard attention, see Appendix A.1.

### 3.3. *Single-layer attention*

Our *single-layer attention* model consists of a single Transformer encoder layer: an attention layer followed by a dense layer, with layer normalization[32] to aid in optimization. Transformer implementations typically use a sine/cosine positional encoding[15] or learned Gaussian positional encoding,[33] rather than the one-hot positional encoding used in our single-layer models.

**Self-Supervised Losses.** Given an MSA, many standard methods estimate Potts model parameters through pseudolikelihood maximization.[2,31] On the other hand, BERT-like attention-based models are typically trained with variants of masked language modeling.[8] Pseudolikelihood is challenging to compute efficiently for generic models, unlike the masked language modeling loss. Both of these losses require computing conditionals of the form $p_\theta(x_i | x_{\setminus M})$, where $M$ is a subset of $\{1, \ldots, L\}$ containing $i$. The losses $\mathcal{L}_{PL}$ and $\mathcal{L}_{MLM}$ for pseudolikelihood and masked language modeling, respectively, are

$$\mathcal{L}_{PL}(\theta; x) = \sum_{i=1}^{L} \log p_\theta(x_i | x_{\setminus i}), \qquad \mathcal{L}_{MLM}(\theta; x, M) = \sum_{i \in M} \log p_\theta(x_i | x_{\setminus M}).$$

Regularization for Potts and factored attention are both based on MRF edge parameters, while single-layer attention is penalized using weight decay. More details can be found in Appendix A.2.

### 3.4. *Pretraining on Sequence Databases*

All single-layer models are trained on a set of evolutionarily related sequences. Given a large database of protein sequences such as UniRef100[34] or BFD,[35,36] these models cannot be trained until significant preprocessing has been done: clustering, dereplication of highly related sequences, and alignment to generate an MSA for each cluster. In contrast, the self-supervised approach taken by works such as Refs. 10–13 applies BERT-style pretraining directly on the database of proteins with minimal preprocessing.

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Given a new sequence of interest and a database of sequences, single-family models require more steps for inference than pretrained Transformers. To apply a single-family model, one must query the database for related sequences, dereplicate the set, align sequences into an MSA, then train a model to learn contacts. On the other hand, a Transformer pretrained on the database simply computes a forward pass for the sequence of interest and its attention activations are used to predict contacts. No explicit querying or aligning is performed.

### 3.5. *Extracting Contacts*

**Potts.** We follow standard practice and extract a contact map $\widehat{C} \in \mathbb{R}^{L \times L}$ from the order-4 interaction tensor $W$ by setting $\widehat{C}_{ij} = \|W^{ij}\|_F$.

**Factored Attention.** Since factored attention is a pairwise MRF, we can compute its order-4 interaction tensor $W$ and use the same procedure as Potts. See Equation A.2.

**Single-Layer Attention.** To produce contacts for an MSA, we compute attention maps from *only* the positional encoding (without sequence) and average attention maps from all heads. Each single-layer attention model is trained on one MSA, so the positional encoding is a feature shared by all sequences in the MSA.

**ProtBERT-BFD.** We extract contacts from ProtBERT by averaging a subset of attention maps for an input sequence $x$. Of the 16 heads in 30 layers, we selected six whose attention maps had the top individual contact precisions over 500 families randomly selected from the Yang *et al.*[6] dataset. Predicted contacts for $x$ are given by averaging the $L \times L$ attention maps from these six heads, then symmetrizing additively. See Appendix Table A1.

**Average Product Correction (APC).** Empirically, Potts models trained with Frobenius norm regularization have artifacts in the outputs $\widehat{C}$. These are removed with the Average Product Correction (APC).[37] Unless otherwise stated, we apply APC to all extracted contacts.

## 4. Results

**Experimental Setup.** We use a set of 748 protein families from Ref. 6 to evaluate all models. For Potts models and single attention layers, we train separate models on each individual MSA. ProtBERT-BFD is frozen for all experiments. We train models using PyTorch-Lightning[38] and Weights and Biases.[39] We extract contacts from each model following the procedure outlined in Appendix A.4.2. We compare predicted contact maps $\widehat{C}$ to true contact maps $C$ using standard metrics based on precision. A particularly important metric is *precision at L*, where $L$ is the length of the sequence.[40,41] This is computed by masking $\widehat{C}$ to only consider positions $\geq 6$ apart, predicting the top $L$ entries to be contacts, and computing precision. We provide more information on data and metrics in Appendix A.4 and on model hyperparameters in Appendix A.5.

**Attention assumptions reflected in 15,051 protein structures.** We examine all 15,051 structures in the dataset in Ref. 6 for evidence of two key properties useful for single-layer attention models: few contacts per residue and the number of contacts scaling linearly in length. In Appendix Figure A2, we see that 80% of the 3,747,101 million residues in these structures have 4 or fewer contacts. Only 1.8% of residues have more than ten contacts. This shows that the row-wise softmax, which encourages each residue to attend to only a few other

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

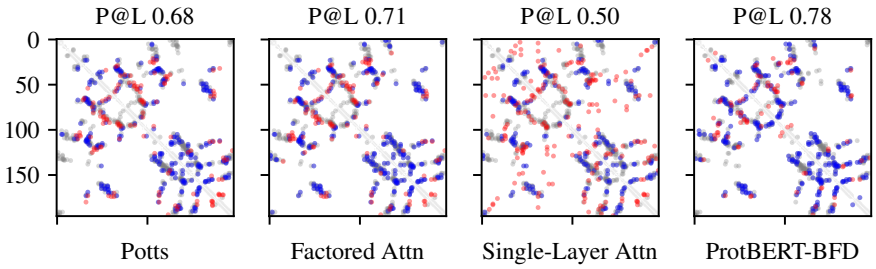
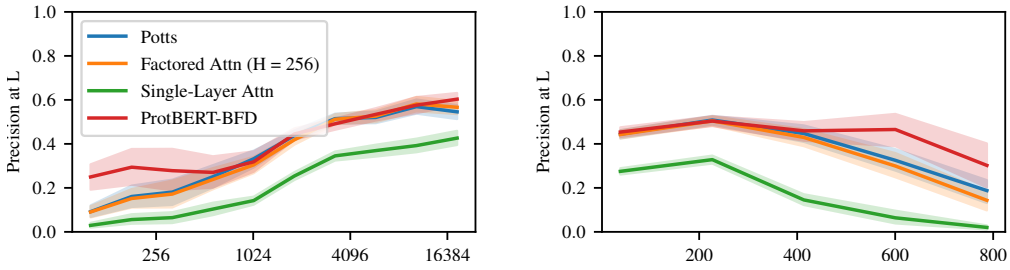
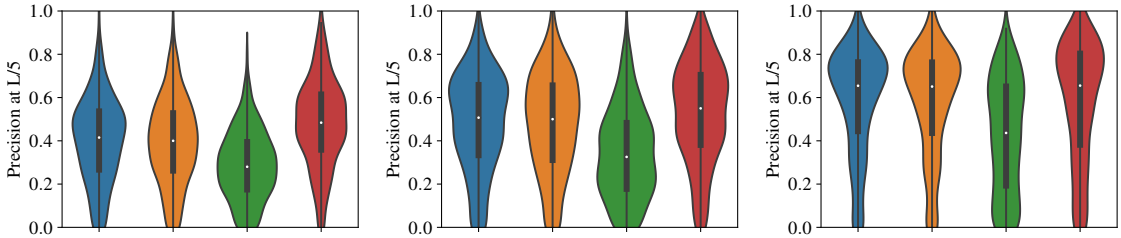



Fig. 1: Predicted contact maps and Precision at $L$ for each model on PDB entry *2BFW*. Blue indicates a true positive, red indicates a false positive, and grey indicates a false negative.




(a) MSA depth (b) Reference length

Fig. 2: Model performance evaluated on MSA depth and reference length. ProtBERT-BFD has higher precision on MSAs with fewer than 256 sequences. For larger MSAs, Potts, Factored Attention, and ProtBERT-BFD perform comparably. Across a variety of protein lengths, Factored Attention performs comparably to Potts with substantially fewer parameters.



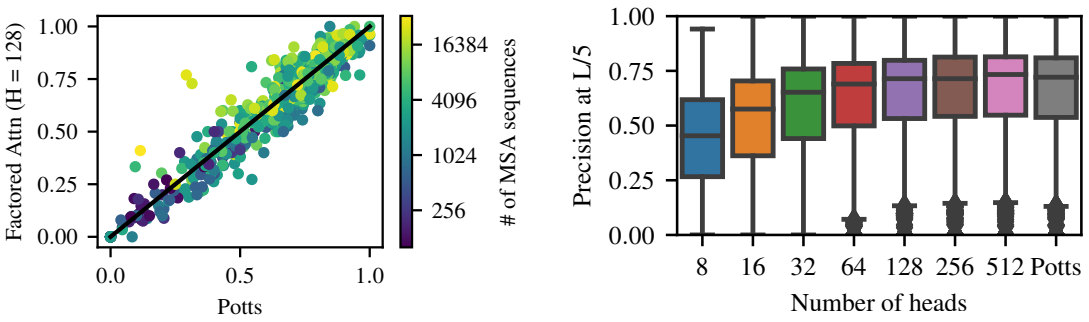

(a) Short-range (b) Medium-range (c) Long-range

Fig. 3: Contact precision for all models stratified by the range of the interaction, with the same color correspondence as in Figure 2a. Potts, Factored Attention, and ProtBERT-BFD perform comparably for long and medium-range contacts, while ProtBERT-BFD has slightly better precision on short-range contacts.

residues per-head, reflects structure found in the data.

**Factored attention matches Potts performance on 748 families.** Figure 1 shows a representative sample of good quality contact maps extracted from all models. Figure 2a summarizes the performance of all models over the set of 748 protein families. Factored at-

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.




(a) 128 heads versus Potts P@$L/5$

(b) P@$L/5$ for varying number of heads.

Fig. 4: Examining impact of number of heads on precision at $L/5$. Left: Comparing performance of Potts and 128 heads over each family shows comparable performance. Right: Precision at $L/5$ drops off slowly until 32 heads, then steeply declines beyond that.

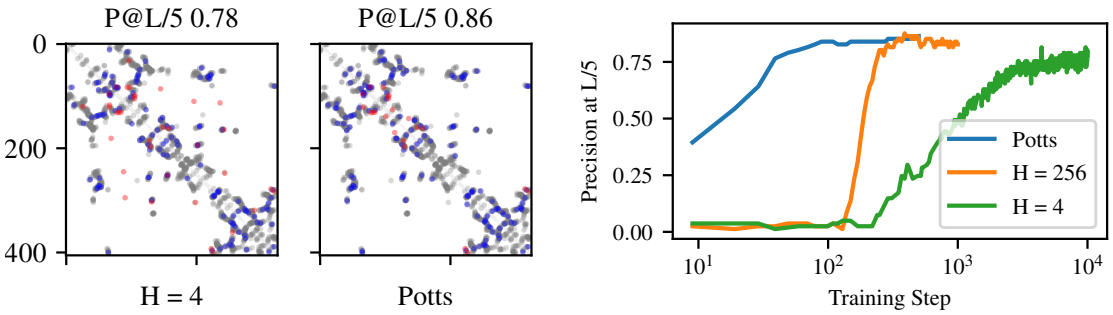
tention, Potts, and ProtBERT-BFD have comparable overall performance, with median precision at $L$ of 0.46, 0.47, and 0.48, respectively. Stratifying by number of sequences reveals that ProtBERT-BFD has higher precision on MSAs with fewer than 256 sequences. For MSAs with greater than 1024 sequences, Potts, factored attention, and ProtBERT-BFD have comparable performance. Single-layer attention is uniformly worse over all MSA depths.

Next, we evaluate the impact of sequence length on performance. Figure 2b shows that factored attention and Potts achieve similar precision at $L$ over the whole range of family lengths, despite factored attention having far fewer parameters for long families. This shows that factored attention can successfully leverage sparsity assumptions where they are most useful.

Long-range contacts are particularly important for downstream structure-prediction algorithms – long-range precision at $L/5$ is reported in both CASP12 and CASP13.[40,41] Figure 3 breaks down contact precisions based on position separation into short ($6 \leq \text{sep} < 12$), medium ($12 \leq \text{sep} < 24$), and long ($24 \leq \text{sep}$). We see that ProtBERT-BFD performs best on short-range contacts, with a median increase of 0.068 precision at $L/5$. On long-range contacts, there is no appreciable difference in performance to Potts and factored attention. Across the range of contact bins, factored attention and Potts perform very similarly.

**Fewer heads can match Potts on $L/5$ contacts.** We probe the limits of parameter sharing by lowering the number of heads in factored attention and evaluating whether fewer heads can be used to precisely estimate contacts. Figure 4a shows that 128 heads can be used to estimate $L/5$ contacts as precisely as Potts over the full set of 748 families. In Figure 4b, we see that factored attention with 32 and 64 heads is still able to achieve reasonable overall performance compared to Potts. 32 and 64 heads have precision at $L/5$ at least as high as Potts for 329 and 348 families, respectively. If we wish to recover the top $L$ contacts, 256 heads are required to match Potts across all families, as seen in Appendix Figure A3. Having more heads than 256 does not further increase performance. Intriguingly, Appendix Figure A4 demonstrates that both Spearman and Pearson correlation between the order-4 interaction

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.




(a) Top $L/5$ contacts on *3n2a*. (b) Training dynamics of models on *3n2a*.

Fig. 5: Factored attention with 4 heads can learn the top $L/5$ contacts on PDB *3n2a*.

tensors of factored attention and Potts improve even when increasing to 512 heads. We do not observe the same trends for increasing head size, as shown in Appendix Figure A5

For some families, the number of heads can be reduced even further. We show an example on the MSA built for PDB entry *3n2a*. In Figure 5a, we see that merely 4 heads are required to recover $L/5$ contacts nearly identical to those recovered by Potts. This shows that shared amino acid features and interaction parameters can enable identical performance with a 300× reduction in parameters. The training dynamics of these models are shown in Figure 5b. Both factored attention with 256 heads and Potts converge after roughly 100 gradient steps, whereas factored attention with 4 heads requires nearly 10,000 steps to converge. In Appendix Figure A6, we show that the top $L$ contacts are significantly worse for 4 heads compared to Potts.

**One set of amino acid features can be used for all families.** Thus far we have only examined models that share parameters within single protein families. Since ProtBERT is trained on an entire database, it can leverage feature sharing across families to attain greater parameter efficiency and improved performance on small MSAs.

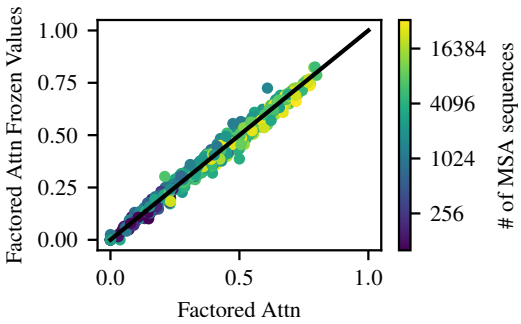



Fig. 6: Precision at $L$ comparison, which illustrates that a single set of frozen value matrices can be used for all families.

To explore the possibility that attention can share parameters across families, we train factored attention using a single set of frozen value matrices. We first train factored attention normally on *3n2a* with 256 heads, then freeze the learned value matrices for the remaining 747 families. The query and key parameters are trained normally. In Figure 6, we compare the precision at $L$ of factored attention with frozen *3n2a* features to that of factored attention trained normally. Using a single frozen set of features results in only 6 families seeing precision at $L$ decrease by more than 0.05, with a

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

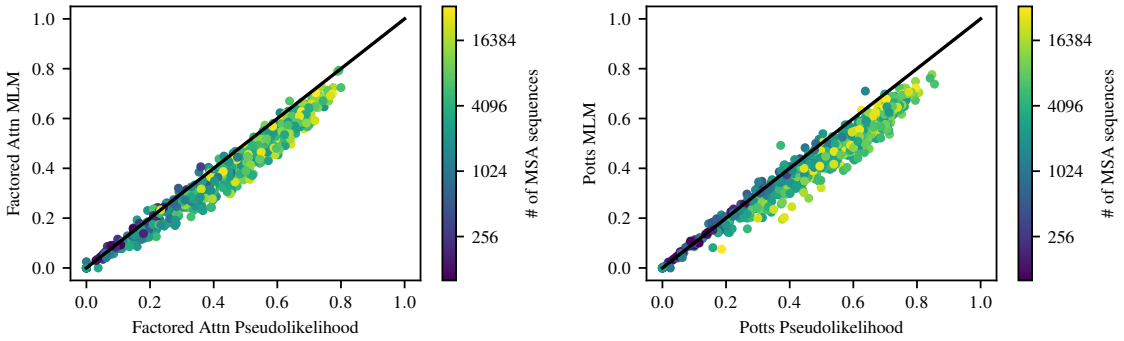



Fig. 7: Effect of loss on precision at $L$ over many families. Pseudolikelihood has a uniform but small benefit over masked language modeling for both models.

maximum drop of 0.11. This suggests that, even for a single-layer model, a single set of value matrices can capture amino acid features across functionally and structurally distinct protein families.

**Factored attention reduces total parameters estimated.** For an MSA of length $L$ with alphabet size $A$, Potts models require $\binom{L}{2}A^2$ parameters. Factored attention with $H$ heads and head size $d$ requires $H(2Ld + A^2)$ parameters. In Figure A7, we plot number of parameters versus length for various values of $H$ and $d = 32$. Potts requires a total of 12 billion parameters to model all 748 families. Factored attention with 256 heads and head size 32 has 3.2 billion parameters; lowering to 128 heads reduces this to 790 million. Half of this reduction comes from 107 families of length greater than 400. ProtBERT-BFD is the most efficient, with 420 million parameters.

**Impact of training loss function.** The choice of loss function had a uniform but small impact for factored attention and Potts. As seen in Figure 7, pseudolikelihood training slightly improves contact accuracy over masked language modeling training.

**Ablations.** APC has a considerable impact on both Potts and factored attention, creating a median increase in precision at $L$ of 0.1 and 0.07, respectively. The effect of APC is negligible for single-layer attention and ProtBERT. Addition of the single-site potential $b_i$ increases performance slightly for attention layers, but not enough to change overall trends. To compare to ProtBERT-BFD, we train our single-layer attention models on unaligned families and found that performance degrades significantly. See Appendix Figures A8-A10.

## 5. Discussion

We have shown that single-layer factored attention models and the ProtBert-BFD Transformer achieve performance comparable to Potts models on unsupervised contact extraction. We have also shown that the assumptions encoded by attention reflect important properties of protein families. These results suggest that attention has a natural role in protein representation learning, without analogy to attention's success in the domain of NLP.

Our results also show that hierarchical signal within and across families can be captured by even simple attention models. The MSA Transformer[42] explicitly ties weights within families to

achieve improved results on contact extraction, showing that modeling of hierarchical structure is beneficial for larger models trained on entire databases. There have been extensive efforts to organize the relationships between protein families and folds, most notably the SCOP[43] and CATH[44] hierarchies. Further leveraging such rich structure will be essential to the development of powerful protein representations.

## Acknowledgements

This research is supported in part by an NIH grant R35-GM134922. Y.S.S. is a Chan Zuckerberg Biohub Investigator. Computing resources were provided by a grant from AWS Cloud Credits for Research.

## References

1. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee and C. J. Langmead, Learning generative models for protein fold families, *Proteins: Structure, Function, and Bioinformatics* **79**, 1061 (apr 2011).
2. M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt and E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer potts models, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **87**, p. 012707 (January 2013).
3. S. Seemayer, M. Gruber and J. Söding, CCMpred - Fast and precise prediction of protein residue-residue contacts from correlated mutations, *Bioinformatics* **30**, 3128 (may 2014).
4. B. Monastyrskyy, D. D'Andrea, K. Fidelis, A. Tramontano and A. Kryshtafovych, New encouraging developments in contact prediction: Assessment of the casp 11 results, *Proteins: Structure, Function, and Bioinformatics* **84**, 131 (2016).
5. S. Wang, S. Sun, Z. Li, R. Zhang and J. Xu, Accurate de novo prediction of protein contact map by ultra-deep learning model, *PLOS Computational Biology* **13**, 1 (01 2017).
6. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov and D. Baker, Improved protein structure prediction using predicted inter-residue orientations, *bioRxiv* , p. 846279 (2019).
7. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu and D. Hassabis, Improved protein structure prediction using potentials from deep learning, *Nature* **577**, 706 (jan 2020).
8. J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding (October 2018).
9. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language models are Few-Shot learners (May 2020).
10. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, D. Guo, M. Ott, C. Lawrence Zitnick, J. Ma and R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences (August, 2020).
11. A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, D. Bhowmik and B. Rost, ProtTrans: Towards cracking the language of life's code through Self-Supervised deep learning and high performance computing (July 2020).
12. R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, P. Chen, J. Canny, P. Abbeel and Y. Song, Evaluating protein transfer learning with TAPE, in *Advances in Neural Information Processing Systems 32*, (Curran Associates, Inc., 2019) pp. 9689–9701.

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

13. A. Madani, B. McCann, N. Naik, N. S. Keskar and others, ProGen: Language modeling for protein generation, *arXiv preprint arXiv* (2020).
14. A. Nambiar, S. Liu, M. Hopkins, M. Heflin, S. Maslov and A. Ritz, Transforming the language of life: Transformer neural networks for protein prediction tasks (June, 2020).
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (Curran Associates, Inc., 2017) pp. 5998–6008.
16. A. X. Lu, H. Zhang, M. Ghassemi and A. Moses, Self-supervised contrastive learning of protein representations by mutual information maximization, *bioRxiv* (2020).
17. A. Shanehsazzadeh, D. Belanger and D. Dohan, Is transfer learning necessary for protein landscape prediction? (2020).
18. C. Hsu, H. Nisonoff, C. Fannjiang and J. Listgarten, Combining evolutionary and assay-labelled data for protein fitness prediction, *bioRxiv* (2021).
19. C. Dallago, J. Mou, K. E. Johnston, B. Wittmann, N. Bhattacharya, S. Goldman, A. Madani and K. K. Yang, Flip: Benchmark tasks in fitness landscape inference for proteins (2021).
20. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, Highly accurate protein structure prediction with alphafold, *Nature* **596**, 583 (2021).
21. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network, *Science* **373**, 871 (2021).
22. L. S. Johnson, S. R. Eddy and E. Portugaly, Hidden Markov model speed heuristic and iterative HMM search procedure, *BMC Bioinformatics* **11**, p. 431 (aug 2010).
23. M. Remmert, A. Biegert, A. Hauser and J. Söding, HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment, *Nature Methods* **9**, 173 (feb 2012).
24. S. W. Lockless and R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families, *Science* **286**, 295 (oct 1999).
25. A. A. Fodor and R. W. Aldrich, On Evolutionary Conservation of Thermodynamic Coupling in Proteins, *Journal of Biological Chemistry* **279**, 19046 (apr 2004).
26. J. Thomas, N. Ramakrishnan and C. Bailey-Kellogg, Graphical models of residue coupling in protein families, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5**, 183 (2008).
27. M. Weigt, R. A. White, H. Szurmant and others, Identification of direct residue contacts in protein–protein interaction by message passing, *Proceedings of the* (2009).
28. S. Burley and G. Petsko, Weakly polar interactions in proteins, *Advances in protein chemistry* **39**, 125 (1988).
29. R. Jaenicke, Stability and stabilization of globular proteins in solution, *Journal of Biotechnology* **79**, 193 (2000).
30. W. R. Taylor and M. I. Sadowski, Structural constraints on the covariance matrix derived from multiple aligned protein sequences, *PLoS One* **6**, p. e28265 (2011).
31. H. Kamisetty, S. Ovchinnikov and D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era, *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15674 (September 2013).
32. J. L. Ba, J. R. Kiros and G. E. Hinton, Layer normalization, *arXiv preprint arXiv:1607.06450* (2016).
33. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, Transformers: State-of-the-art natural

Biocomputing 2022 Downloaded from www.worldscientific.com
by COLD SPRING HARBOR LABORATORY on 03/31/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

language processing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online, 2020).
34. B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder and C. H. Wu, UniRef: comprehensive and non-redundant UniProt reference clusters, *Bioinformatics* **23**, 1282 (May 2007).
35. M. Steinegger and J. Söding, Clustering huge protein sequence sets in linear time, *Nature communications* **9**, 1 (2018).
36. M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger and J. Söding, HH-suite3 for fast remote homology detection and deep protein annotation, *BMC Bioinformatics* **20**, p. 473 (sep 2019).
37. 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 (February 2008).
38. W. Falcon, Pytorch lightning, *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning* **3** (2019).
39. L. Biewald, Experiment tracking with weights and biases (2020), Software available from wandb.com.
40. J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych and A. M. Bonvin, Assessment of contact predictions in casp12: co-evolution and deep learning coming of age, *Proteins: Structure, Function, and Bioinformatics* **86**, 51 (2018).
41. R. Shrestha, E. Fajardo, N. Gil, K. Fidelis, A. Kryshtafovych, B. Monastyrskyy and A. Fiser, Assessing the accuracy of contact predictions in casp13, *Proteins: Structure, Function, and Bioinformatics* **87**, 1058 (2019).
42. R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu and A. Rives, Msa transformer, in *Proceedings of the 38th International Conference on Machine Learning*, eds. M. Meila and T. Zhang, Proceedings of Machine Learning Research, Vol. 139 (PMLR, 18–24 Jul 2021).
43. J.-M. Chandonia, N. K. Fox and S. E. Brenner, Scope: classification of large macromolecular structures in the structural classification of proteins—extended database, *Nucleic acids research* **47**, D475 (2019).
44. I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, C. S. Pang, L. Woodridge, C. Rauer, N. Sen *et al.*, Cath: increased structural coverage of functional space, *Nucleic acids research* **49**, D266 (2021).
45. D. P. Brown, N. Krishnamurthy and K. Sjölander, Automated protein subfamily identification and classification, *PLoS Comput Biol* **3**, p. e160 (2007).
46. D. Malinverni and A. Barducci, Coevolutionary analysis of protein subfamilies by sequence reweighting, *Entropy* **21**, p. 1127 (2019).
47. S. Ovchinnikov, H. Kamisetty and D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information, *Elife* **3**, p. e02030 (May 2014).
48. J. Dauparas, H. Wang, A. Swartz, P. Koo, M. Nitzan and S. Ovchinnikov, Unified framework for modeling multivariate distributions in biological sequences (June 2019).
49. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu and U. Consortium, Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics* **31**, 926 (2015).