Appendix A.

A.1. *Recovering Factored Attention from Standard Attention*

Potts and Factored Attention estimate a single undirected graphical model from the training data. While a single graph can be a good approximation for the structure associated with a protein family, many families have *subfamilies* with different functional specializations and even different underlying contacts.[45,46] Since subfamily identity is rarely known, allowing edge weights to be a function of sequence could enable the estimation of a family of graphs.

In the language of the Transformer, factored attention estimates a single graph because it computes queries and keys using only the positional encoding. We show more precisely that factored attention can be recovered from standard attention by computing queries and keys from one-hot positional encodings and values from one-hot sequence embeddings.

Single attention layer. Given a sequence of dense vectors $x = (x_1, \ldots, x_L)$ with $x_i \in \mathbb{R}^p$, the attention mechanism of the Transformer encoder (multihead scaled dot product self-attention) produces a continuous representation $y \in \mathbb{R}^{L \times p}$. If head size is d, this representation is computed using H heads (W_Q, W_K, W_V), where $W_Q, W_K, W_V \in \mathbb{R}^{p \times d}$. Queries, keys, and values are defined as $Q = xW_Q, K = xW_K, V = xW_V$. For a single head (W_Q, W_K, W_V), the output is given by

$$y = \text{softmax}\Big(\frac{QK^T}{\sqrt{d}}\Big)V.$$

The full output in $\mathbb{R}^{dH}$ is produced by concatenating all head outputs. A single Transformer encoder layer passes the output through a dense layer, applying layer-norms and residual connection to aid optimization.

For the first layer, the input x is a sequence of discrete tokens. To produce a dense vector combining sequence and position information, positional encodings and sequence embeddings are combined. The positional encoding $E_{pos} \in \mathbb{R}^{L \times e}$ produces a dense vector of dimension e for each position i. The sequence embedding $E_{seq} \in \mathbb{R}^{A \times e}$ maps each element of the vocabulary to a dense vector of dimension e. Typically these are combined through summation to produce a dense vector $\tilde{x}_i = E_{seq}(x_i) + E_{pos}(i)$, which is input to the Transformer as described above.

For this paper, we use only multi-head self-attention without the dense layer, layer norm, or residual connections, as these drastically hurt performance when employed for one layer.

Factored attention from standard attention. Written explicitly, the input Transformer layer computes queries for a single head with $Q = (E_{pos} + E_{seq}(x))\, W_Q$. Keys and values are computed similarly. To recover factored attention, we instead compute queries and keys via $Q = E_{pos}W_Q$ and $K = E_{pos}W_K$, while values are given by $V = E_{seq}(x)W_V$. For simplicity, we one-hot encode both position and sequence, which corresponds using identity matrices $E_{pos} = I_L \in \mathbb{R}^{L \times L}$ and $E_{seq} = I_A \in \mathbb{R}^{A \times A}$.

Implicit single-site term in single-layer attention. For a single layer of attention, the product $E_{pos}W_V$ is a matrix in $\mathbb{R}^{L \times A}$. This matrix does not depend on sequence inputs, thus allowing it to act as a single-site term. This suggests why inclusion of an explicit single-site term in Figure A10 had no effect for single-layer attention.

A.2. *Losses*

The loss for all three models is of the form

$$\ell(\theta; x) = \mathcal{L}(\theta; x) + cR(\theta), \tag{A.1}$$

where $\mathcal{L}$ is either pseudolikelihood or masked language modeling and R is a regularizer.

Potts regularization. Consider the order-4 interaction tensor W, where $W^{ij} \in \mathbb{R}^{A \times A}$ gives the parameters associated to edge (i, j). We regularize W by setting $R(\theta) = \sum_{i<j} \|W^{ij}\|_F^2$. This term is multiplied by $\lambda \cdot L \cdot A$, following.[47]

Factored attention regularization. Since factored attention is also a fully connected pairwise MRF, we use identical regularization to that of Potts. The order-4 tensor W is given by

$$W_{ab}^{ij} = \sum_{h=1}^{H} \operatorname{symm}\Big(\operatorname{softmax}\left(W_Q^h W_K^{h\,T}\right)\Big)_{ij} W_V^h(a, b). \tag{A.2}$$

Single-layer attention regularization. Due the lack of an MRF interpretation for single-layer attention, we chose to use weight decay as is typically done for attention models. This corresponds to setting $R(\theta)$ to be the sum of Frobenius norm squared for all weights W_Q, W_K and W_V.

Single-site term. When any model has a single-site term, we follow standard practice and regularize its Frobenius norm squared.

A.3. *ProtBERT-BFD head selection*

layer	head	P@L
29	7	0.517
29	8	0.396
29	4	0.394
29	2	0.353
29	11	0.333
29	0	0.299
28	3	0.275
29	15	0.177
29	6	0.167
29	12	0.158
28	4	0.141
29	9	0.139
28	6	0.125
28	5	0.125
3	4	0.115
28	11	0.106

Table A1: The top 16 heads in ProtBERT-BFD whose attention maps gave the most precise contact maps across 500 validation families. Most of the top performing heads are found in the last layer. The top six heads were selected for our contact extraction in all results.

A.4. *Data and Metrics*

A.4.1. *Selection of Protein Families*

We used the following sets of families for model development:

(1) A set of 748 families was chosen for performance evaluation. All metrics reported in the paper are on this set, with a single choice of hyperparameters for Potts models, factored attention, and standard attention. The 748 families were chosen randomly from the Yang *et al.*[6] dataset, which consists of 15,051 MSAs generated from the databases UniClust30 and UniRef100,[34] as well as metagenomic datasets. Our random sample is representative of the range of MSA depths and protein lengths, see Figure A1.

(2) A set of six families from the 748 was chosen to choose hyperparameters for single-layer attention. They were chosen to span a range of MSA depth (large and small), as well as three different regimes of Potts performance (Good, Ok, Poor). These families were used to tune hyperparameters as described in Section A.5.1. See Table A2.

(3) A set of ten families from the 748 was chosen where factored attention performed very poorly in our initial experiments. Half were chosen to be long proteins and the other half to be short. This set was used to optimize learning rate and regularization for factored attention to ensure reasonable model performance. See Table A3.

(4) 500 entirely disjoint families were further selected randomly from[6] and used to compute average precision for each head in ProtBERT-BFD.[11] Performance on these families was used for selecting the top 6 heads, see Table A1.

PDB	Sequences	Length	Potts Performance
3er7_1_A	33673	118	Good
5fo5_1_B	17560	88	Ok
2w18_1_A	33619	308	Poor
4gnr_1_A	2073	351	Good
5mkc_1_A	515	207	Ok
3e9l_1_A	146	292	Poor

Table A2: 6 families chosen for hyperparameter optimization for single-layer attention.

PDB	Sequences	Length
4k61_1_A	2145	140
4l3r_1_A	5535	143
3cy4_1_B	1064	154
6fdg_1_A	2325	155
3p6b_1_B	4353	186
1jm1_1_A	17130	202
4yt2_1_A	15481	343
3vmm_1_A	4383	471
4egc_1_A	9929	539
3gq7_1_A	6568	605

Table A3: 10 families chosen for hyperparameter optimization for factored attention

Fig. A1: The length and MSA size distribution for our 748 family subset (red) compared to the full 15,051 families in the trRosetta dataset selected for training

A.4.2. *Producing Contact Maps*

A PDB structure gives 3D coordinates for every atom in a structure. We use Euclidean distance between the beta carbons to define distance between any pair of positions. In the case of glycine, the alpha carbon is used. A pair of positions where this distance is less than $8\mathring{A}$ is declared to be a contact.

A.4.3. *Scoring Contact Predictions*

Given a predicted contact map $\widehat{C} \in \mathbb{R}^{L \times L}$ and a true contact map $C \in \{0,1\}^{L \times L}$, we describe metrics for scoring $\widehat{C}$.

A sequence $x = (x_1, \ldots, x_L)$ of length L has $\binom{L}{2}$ potential contacts. Since we see $\mathcal{O}(L)$ contacts, contact prediction a sparse prediction task. Accordingly, we focus on precision-recall based quantitative analyses of $\hat{C}$. Common practice in the field is to sort all $\binom{L}{2}$ entries of $\hat{C}$ in decreasing order and evaluate precision at various length thresholds, such as the top L or $L/10$ predictions.[41] Note that this analysis is similar to choosing recall cutoffs along a precision-recall curve, where sorted length index plays the role of recall on the x axis. Unlike

recall, length-based cutoffs do not rely on knowledge of the actual number of contacts. In addition to the precision at various length (recall) cutoffs, we also computed Area Under the Precision-Recall Curve (AUC), which we define as the average of Precision @ L for $L = 1, ..., 10$. AUC is a widely used metric for comparing classifiers when the positive class is rare.

A.5. *Hyperparameters*

Potts. We used $\lambda = 0.5$, learning rate of 0.5, and batch size 4096. Pad, gap, and mask were all encoded with the same token. The Potts model is trained using a modified version of Adam presented in Ref. 48. This modification was made to improve performance of Adam to match that of L-BFGS.

Factored attention. We AdamW with a learning rate of 5×10^{-3} and set $\lambda = 0.01$. The default head size was set to 32 unless stated otherwise.

Single-layer attention. We set embedding size of 256, head size of 64, and number of heads 128. The model is trained with AdamW using a learning rate of 5×10^{-3} and weight decay of 2×10^{-3}. Attention dropout of 0.05 is also applied. The batch size is 32 and mask prob for masked language modeling is 0.15. We use a separate mask token and pad,gap token.

ProtBERT-BFD. ProtBERT-BFD has 30 layers each with 16 heads and a hidden size of 1024. The training dataset is a mixture of UniRef50[49] and BFD. It has 2,122 million protein sequences. See[11] more information.

A.5.1. *Hyperparameter Sweep Details*

Potts. The Potts model implementation using psuedolikelihood has been optimized by others, so we did not tune performance. Since performance with MLM was comparable to pseudolikelihood, we did not sweep for MLM either.

Single-layer attention. Standard attention is by far the most sensitive model to hyperparameters. To find a reasonable set of hyperparameters, we first swept over the six families in Table A2, performing a grid search over

- $H \in \{32, 64, 128, 256, 512\}$
- $d \in \{32, 64, 128, 256, 512\}$
- $e \in \{128, 256, 512\}$
- attention dropout in $\{0, 0.05, 0.1\}$
- learning rate in $\{5 \times 10^{-3}, 1 \times 10^{-2}\}$
- weight decay in $\{0, 1 \times 10^{-3}, 2 \times 10^{-3}\}$

We found that the choice $H = 256$, $d = 64$, $e = 256$, attention dropout of 0.05, learning rate of 5×10^{-3} and weight decay of $\times 10^{-3}$ performed well across all six families. Due to GPU memory constraints, we had to set $H = 128$ for further runs.

Factored attention. We swept factored attention over the families in Table A3, performing a grid search over

- learning rate in $\{1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}\}$
- regularization coefficient in $\{1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}\}$

We found that learning rate of 5×10^{-3} and regularization of 0.01 were effective, but that other configurations such as regularization of 5×10^{-3} also performed well. Both H and d are evaluated extensively in our results.

A.6. *Additional Figures*

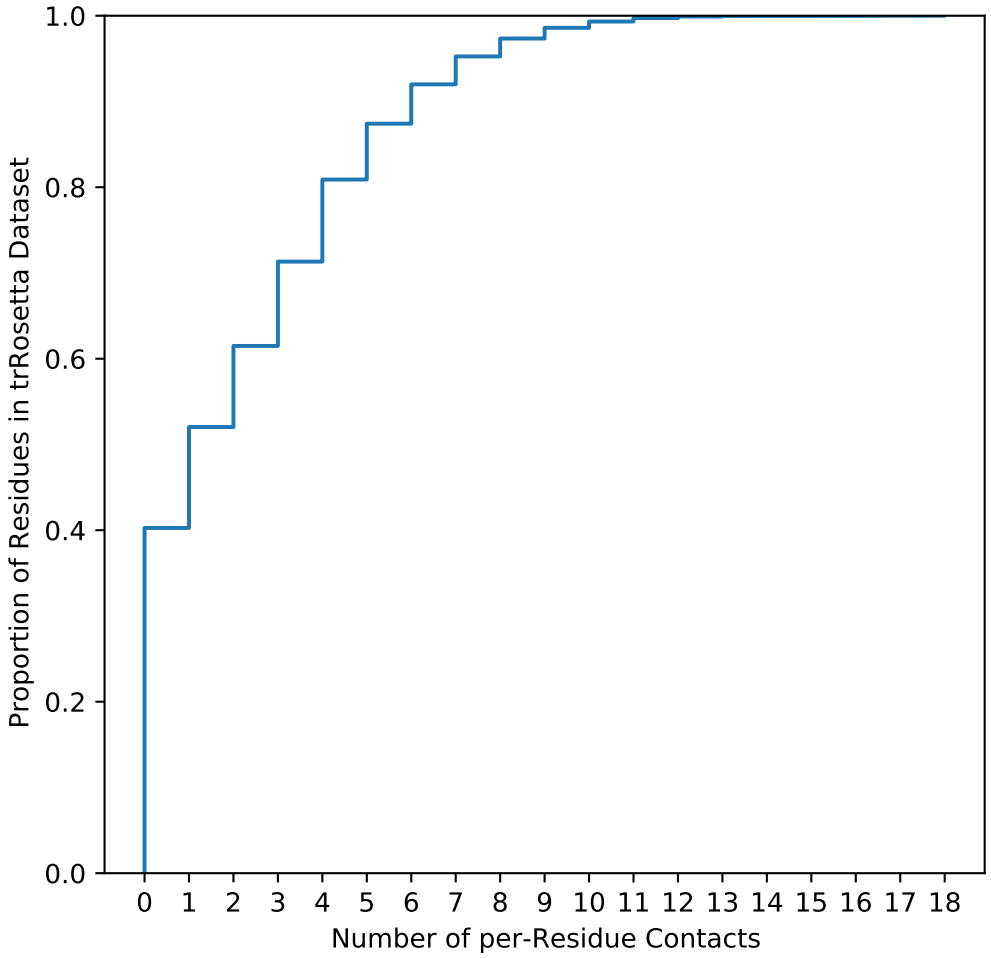

Fig. A2: The empirical CDF of number of per-residue contacts for 3,747,101 residues in 15,051 structures in the trRosetta dataset.

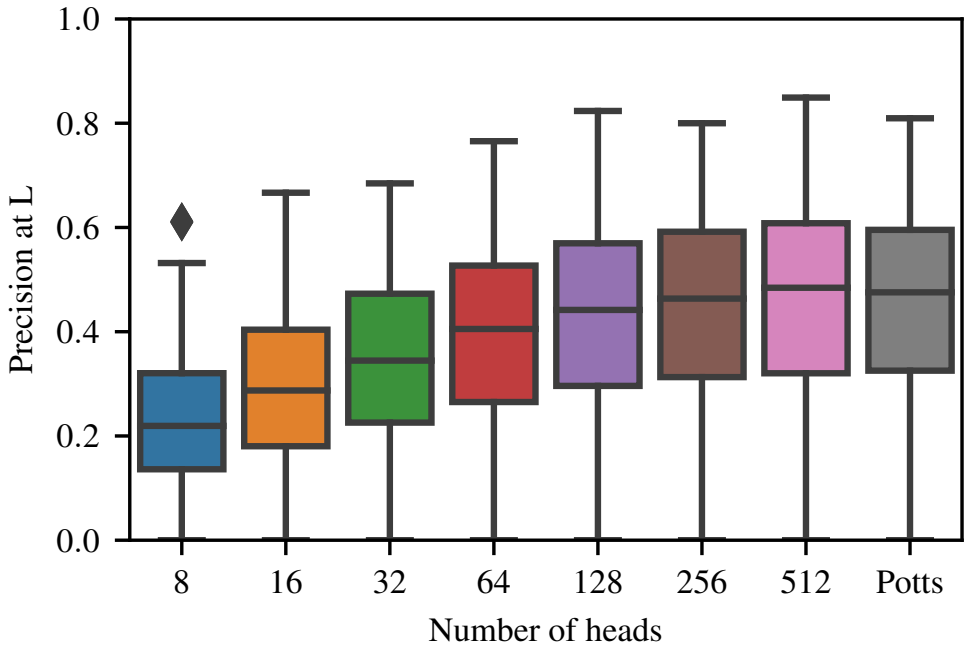

Fig. A3: Reducing the number of heads causes a much steeper decrease in precision at L.

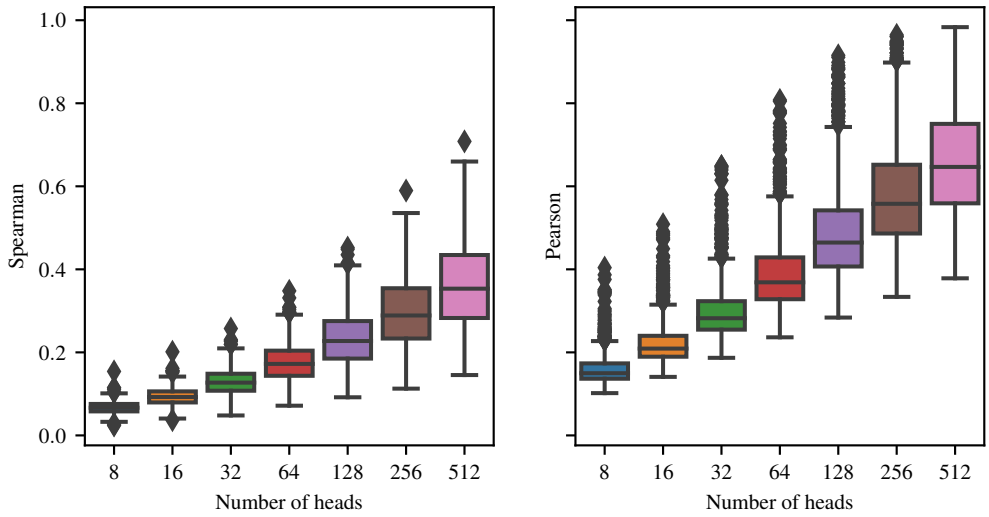

Fig. A4: Effect of number of heads on correlation between the order-4 weight tensors for factored attention (see Equation A.2) and Potts (see Section 3).

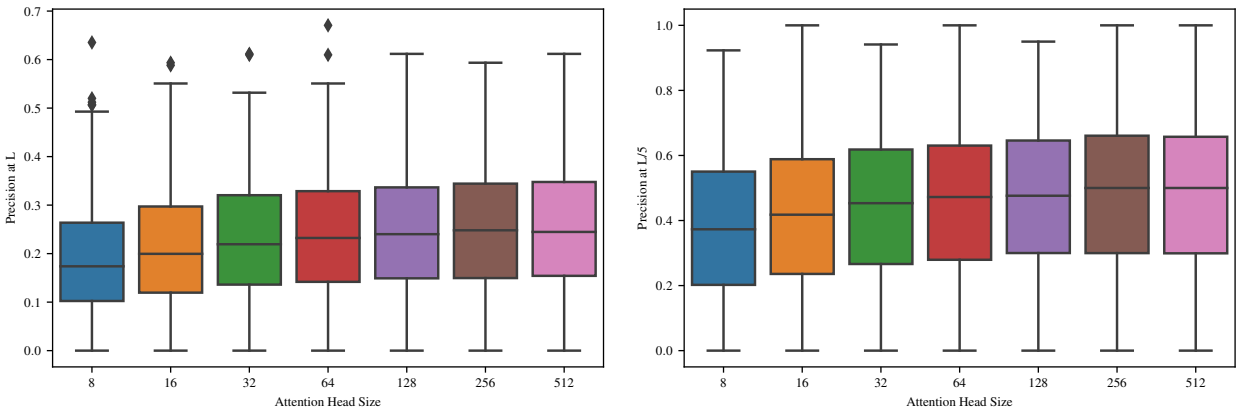

Fig. A5: Effect of head size on factored attention precision at L and $L/5$ over 748 families. Increasing head size has a small effect on precision, though not nearly as pronounced as the number of heads.

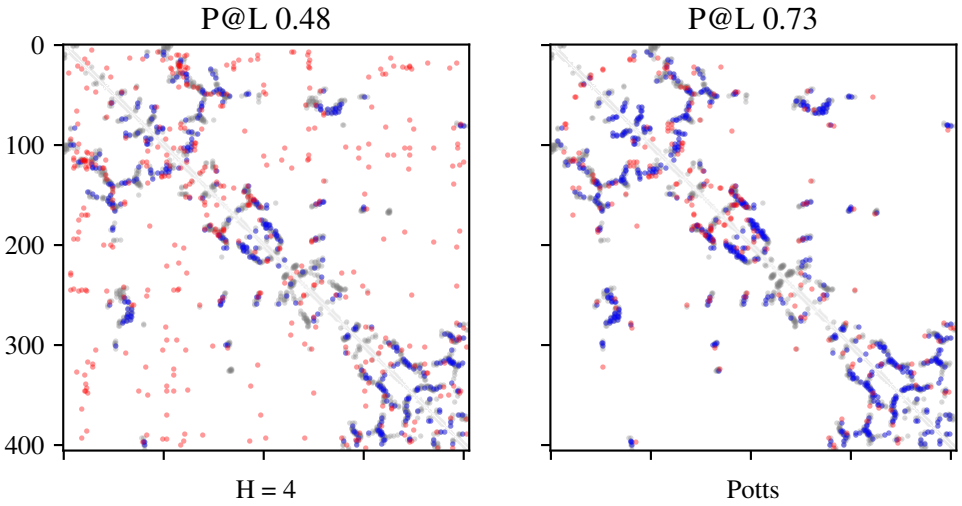

Fig. A6: Factored attention with 4 heads has degraded performance for precision at L for family *3n2a*.

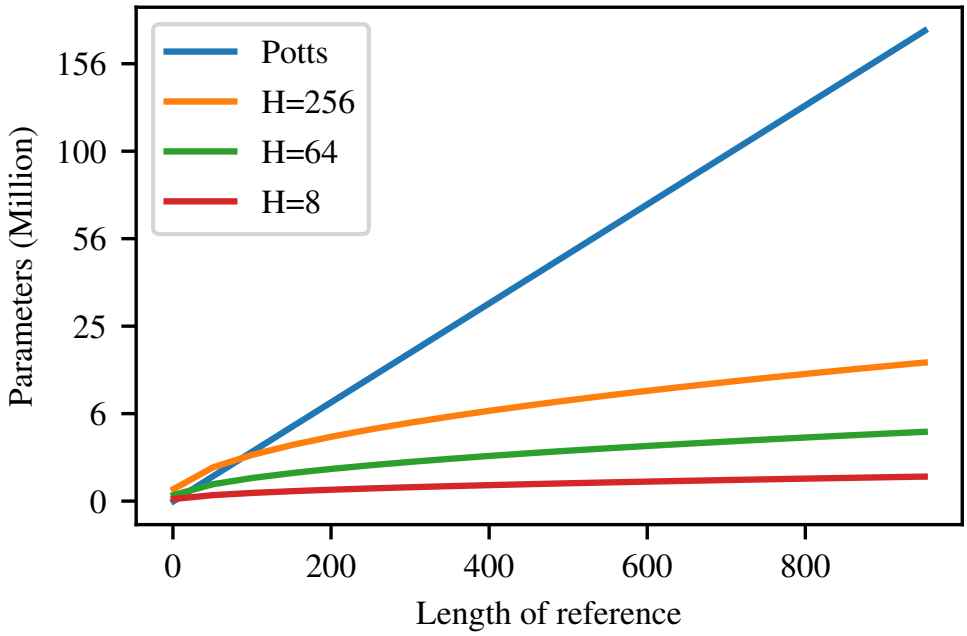

Fig. A7: Number of parameters versus length for MRF models.

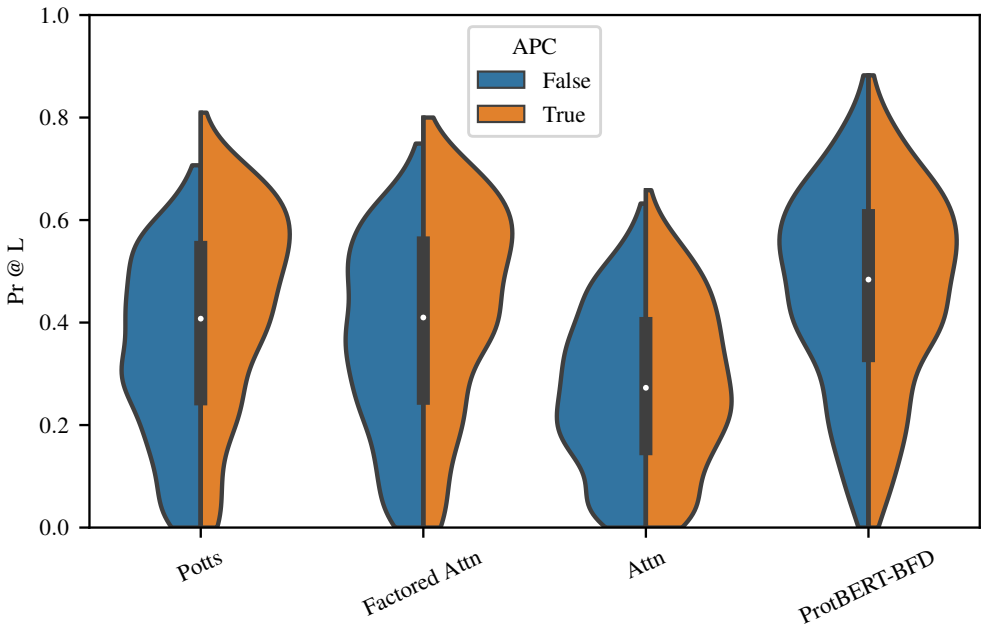

Fig. A8: APC has a significant positive effect on the performance of Potts and factored attention. It makes only a slight difference on the performance of the other two models.

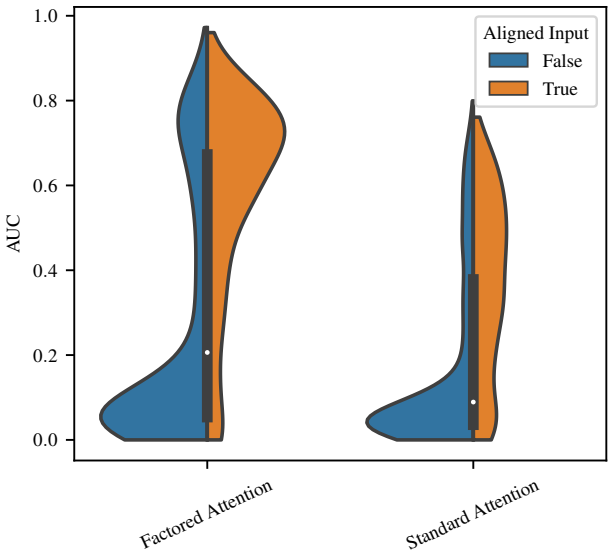

Fig. A9: Training on unaligned families degrades performance on almost all families.

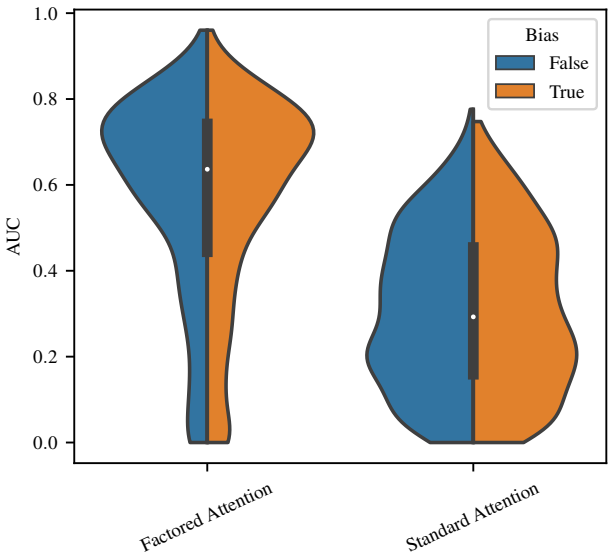

Fig. A10: The addition of a single-site term to either factored or standard attention produces little additional benefit.