BEYOND PAIRWISE ATTENTION: HIGHER-ORDER MODULAR ATTENTION FOR EFFICIENT SEQUENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence modeling tasks can involve intrinsic higher-order dependencies, while standard self-attention assigns scores to token pairs and does not explicitly parameterize such interactions. We introduce **H**igher-**O**rder **M**odular **A**ttention (HOMA), which fuses pairwise attention with an explicit triadic attention pathway made tractable through overlapping blocks, local windows, and a low-rank projection. We compare HOMA with matched pairwise and purely triadic baselines on controlled PARITY and MATCH3 tasks, as well as TAPE benchmarks. HOMA is competitive with or outperforms the baselines, with its clearest advantages when the underlying dependencies extend beyond the explicitly modeled triadic order. The learned nonlinear fusion of these two representations further improves convergence speed and parameter efficiency. Overall, our results provide empirical evidence that HOMA is an effective attention design when task structure extends beyond pairwise interactions.

1 INTRODUCTION

Transformer architectures have become a standard backbone for sequence learning and form the foundation of modern large language models (Vaswani et al., 2017). In standard self-attention, the score between tokens i and j is given by the dot product of their query and key projections, making the scoring mechanism inherently pairwise. While this formulation has been remarkably effective, recent theory has identified settings that expose representational limitations of pairwise attention when modeling higher-order dependencies.

One such limitation concerns how efficiently pairwise attention can communicate joint information across multiple tokens. Sanford et al. (2024) formalize a sharp distinction through two sequence-to-sequence tasks that differ only in arity. In MATCH2, a token must determine whether it participates in a pair whose values sum to zero modulo M; MATCH3 asks the analogous question for a triple. They show that, in the one-layer setting, solving the inherently three-way MATCH3 task with standard pairwise attention requires model resources to grow with sequence length, whereas a single third-order attention unit can solve it with constant embedding dimension. Kozachinskiy et al. (2025) further show that this limitation persists even with arbitrarily high numerical precision, and establish analogous one-layer lower bounds for function composition (Peng et al., 2024) and binary relation composition.

A complementary limitation concerns sensitivity to individual inputs. Hahn (2020) shows that, under fixed-size self-attention, changing a single input symbol changes the encoder output by only $O(1/N)$ as the sequence grows. This creates a difficulty for globally sensitive functions such as PARITY, whose output changes whenever any single bit is flipped. Chiang & Cholak (2022) later showed that PARITY is representable at arbitrary length, but their construction is highly parameter-sensitive and standard Transformers failed to learn it across the settings they tested. Consistent with this, Bhattamishra et al. (2020) report substantial difficulties in learning and length-generalizing PARITY with Transformers. These concerns motivate an empirical question: when a target depends on genuinely higher-order structure, how readily can pairwise attention learn that dependency under finite model and optimization budgets?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Mechanisms that move beyond pairwise attention have appeared in several forms, yet increasing interaction order introduces its own computational trade-offs. The 2-simplicial Transformer (Clift & Murfet, 2020) combines pairwise and triadic attention, but controls the cubic cost of the latter by restricting its partner pair to a small set of learned virtual entities, yielding a global, content-based approximation. Higher-order tensor attention (Sanford et al., 2024) generalizes attention to s-way multilinear interactions, but is introduced primarily as a theoretical construction because dense higher-order attention scales rapidly in cost. Strassen attention (Kozachinskiy et al., 2025) instead scores triples through a structured sum of pairwise inner products that permits sub-cubic evaluation. Hypergraph attention (Bai et al., 2021) captures higher-order relations over an explicitly constructed hypergraph, while poly-attention (Chakrabarti et al., 2026) introduces a general framework for higher-order attention that encompasses standard, tensor, and Strassen attention as special cases. Collectively, these works demonstrate that explicit higher-order interactions are feasible, but typically trade increased interaction order against computation, restricted interaction structure, or both.

Nevertheless, empirical evidence for higher-order attention on natural sequence data remains comparatively limited, with much of the literature centered on theoretical analyses or controlled and synthetic settings (Sanford et al., 2024; Clift & Murfet, 2020; Kozachinskiy et al., 2025). One natural domain in which higher-order attention may offer an advantage is protein sequence modeling, as both protein structure and function exhibit dependencies beyond additive or pairwise effects. For Secondary Structure, β-turns impose cooperative backbone constraints across three or more residues (Venkatachalam, 1968; Smith & Pease, 1980), while helix-terminal stabilization can involve coordinated interactions among several residues (Chakrabartty et al., 1993). The relevance of multi-residue structure is also reflected in AlphaFold2, whose Evoformer repeatedly applies triangle-based operations to residue-pair representations (Jumper et al., 2021). Contact Prediction provides a complementary motivation, as residue covariation may be mediated by other positions rather than reflect direct physical contact (Morcos et al., 2011), making third-residue context potentially useful for distinguishing direct from indirect associations. Finally, at the functional level, epistasis makes the effect of a substitution depend on its genetic background (Weinreich et al., 2013; Poelwijk et al., 2016); additive-plus-pairwise models can therefore be insufficient (Horovitz & Fersht, 1990; Wells, 1990), with substantial three-way and higher-order components observed across genotype–phenotype landscapes (Poelwijk et al., 2019; Sailer & Harms, 2017; Wu et al., 2016).

We introduce HOMA, a Higher-Order Modular Attention mechanism designed as a drop-in replacement for the self-attention sublayer of standard Transformer backbones. Rather than replacing pairwise attention, HOMA augments it with an explicit triadic pathway and combines the two representations through learned fusion. Drawing on efficient pairwise-attention implementations (Beltagy et al., 2020; Zaheer et al., 2020; Shen et al., 2018; Qiu et al., 2020), HOMA's computation is localized through overlapping blocks and local triadic windows, with a low-rank projection limiting additional parameter growth. Our central question is whether making three-way interactions explicit and combining them with pairwise attention yields a more learnable and parameter-efficient attention design under controlled computational budgets, rather than whether pairwise Transformers are fundamentally incapable of representing such dependencies. We evaluate this question against pairwise baselines and a purely triadic variant within a modular pipeline with matched architectures, training protocols, and efficiency profiling. We first use controlled PARITY and MATCH3 experiments to probe interaction order, capacity, and learnability, and then evaluate whether these controlled advantages extend to TAPE Secondary Structure, Contact Prediction, and Fluorescence. Across controlled tasks, HOMA shows its clearest advantages when the underlying dependencies extend beyond pairwise interactions, with gains in accuracy, convergence, and parameter efficiency across several settings. On protein sequence benchmarks, these advantages are task-dependent: HOMA shows its clearest gains on Secondary Structure and Contact Prediction, while remaining competitive on Fluorescence.

2 METHOD

As shown in Figure 1, HOMA augments a standard pairwise attention backbone with a triadic interaction pathway. Let $X \in \mathbb{R}^{L \times d_{\text{model}}}$ denote a length-L sequence of embeddings. We form the standard projections

$$Q = XW^{(Q)}, \qquad K = XW^{(K)}, \qquad V = XW^{(V)},$$

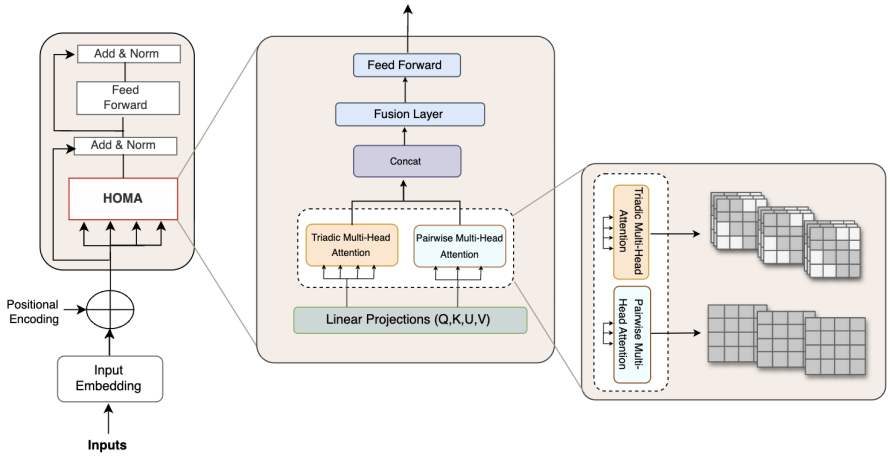


Figure 1: Transformer layer with higher-order modular attention. HOMA replaces the self-attention sublayer and computes pairwise and triadic multi-head attention in parallel from shared projections Q, K, V, and U. Their outputs are concatenated, fused, and mapped through the standard output projection.

and an additional projection for the triadic pathway,

$$U = XW^{(U)},$$

where $W^{(Q)}, W^{(K)}, W^{(V)}, W^{(U)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are learned parameters. We use H attention heads with $d_{\text{head}} = d_{\text{model}}/H$ and present the following computations for a single head. Superscripts (2) and (3) denote pairwise and triadic quantities, respectively.

Pairwise attention. Let $Q, K, V \in \mathbb{R}^{L \times d_{\text{head}}}$, where Q_i denotes the i-th row and Q_{ic} its c-th component. The scaled dot-product scores are

$$S^{(2)}_{ij} = \frac{Q_i K_j^\top}{\sqrt{d_{\text{head}}}}, \qquad S^{(2)} \in \mathbb{R}^{L \times L}.$$

Normalizing across j yields attention weights

$$A^{(2)}_{ij} = \frac{\exp(S^{(2)}_{ij})}{\sum_{j'=1}^{L} \exp(S^{(2)}_{ij'})},$$

and the pairwise output is

$$O^{(2)}_i = \sum_{j=1}^{L} A^{(2)}_{ij} V_j, \qquad O^{(2)} \in \mathbb{R}^{L \times d_{\text{head}}}.$$

Triadic attention. Let $U \in \mathbb{R}^{L \times d_{\text{head}}}$. The triadic pathway assigns weights to ordered pairs (j, z) based on query position i via

$$S^{(3)}_{ijz} = \frac{1}{\sqrt{d_{\text{head}}}} \sum_{c=1}^{d_{\text{head}}} Q_{ic} K_{jc} U_{zc}, \qquad S^{(3)} \in \mathbb{R}^{L \times L \times L}.$$

Normalizing across (j, z) yields

$$A^{(3)}_{ijz} = \frac{\exp(S^{(3)}_{ijz})}{\sum_{j'=1}^{L} \sum_{z'=1}^{L} \exp(S^{(3)}_{ij'z'})}.$$

For each pair (j, z) we form a quadratic value interaction

$$\tilde{V}_{jz} = V_j \odot V_z \in \mathbb{R}^{d_{\text{head}}},$$

where $\odot$ denotes elementwise multiplication. The triadic output is

$$O_i^{(3)} = \sum_{j=1}^{L} \sum_{z=1}^{L} A_{ijz}^{(3)} \tilde{V}_{jz}, \qquad O^{(3)} \in \mathbb{R}^{L \times d_{\text{head}}}.$$

Fusion of pairwise and triadic pathways. For each position i, the pairwise and triadic outputs are concatenated and passed through a fusion network

$$O_i^{\text{cat}} = \left[O_i^{(2)} \,\|\, O_i^{(3)} \right] \in \mathbb{R}^{2d_{\text{head}}}, \qquad O_i^{\text{fuse}} = g\big(O_i^{\text{cat}}\big) \in \mathbb{R}^{d_{\text{head}}},$$

where $g(\cdot)$ is a two-layer MLP with ReLU activation that acts as an adaptive gate, allocating representational capacity between the pairwise and triadic channels as dictated by the data. Before softmax normalization, masks are applied to both the pairwise and triadic attention weights. After concatenating the H head outputs, the standard output projection $W^{(O)} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ is applied.

For comparison, we also consider BLOCKWISE-3D, a purely triadic mechanism that computes $S^{(3)}$, $A^{(3)}$, and $O^{(3)}$ identically to HOMA's triadic pathway; this attention variant omits the pairwise pathway and fusion network and applies $W^{(O)}$ directly to the concatenated triadic head outputs.

2.1 EFFICIENT IMPLEMENTATION

The triadic pathway is the primary computational challenge; naive evaluation of $S_{ijz}^{(3)}$ over all triplets scales as $\mathcal{O}(L^3 d_{\text{head}})$ in compute and $\mathcal{O}(L^3)$ in storage, which is infeasible for typical protein lengths. HOMA and BLOCKWISE-3D address this through three complementary optimizations: overlapping block decomposition to localize computation, windowed triadic attention to reduce cubic scaling within each block, and a low-rank U projection to control parameter growth.

Overlapping block decomposition. Block-granular computation aligns with two complementary properties: contiguous tiling maximizes GPU memory throughput and Tensor Core utilization (Dao et al., 2022), and neighboring keys empirically carry similar attention weights (Jiang et al., 2024), so contiguous blocks capture coherent attention mass without scattering it across distant positions. We partition the sequence into overlapping blocks of length ℓ with stride s, giving

$$T = \left\lfloor \frac{L - \ell}{s} \right\rfloor + 1$$

blocks, and block outputs are merged back to the full sequence by overlap-averaging. In HOMA, both pairwise and triadic attentions are computed independently within each block.

Windowed triadic attention within blocks. Within each block, we restrict (j, z) to a local neighborhood of size $w \ll \ell$ around each query position. This reduces per-block triadic compute from $\mathcal{O}(\ell^3 d_{\text{head}})$ to $\mathcal{O}(\ell w^2 d_{\text{head}})$ while preserving local higher-order interactions.

Low-rank U projection. To limit parameter growth in the triadic pathway, we factorize $W^{(U)}$ as

$$W^{(U)} = W^{(U_u)} W^{(U_v)}, \qquad W^{(U_u)} \in \mathbb{R}^{d_{\text{model}} \times r}, \quad W^{(U_v)} \in \mathbb{R}^{r \times d_{\text{model}}}, \quad r \ll d_{\text{model}}.$$

This retains the expressivity of an additional projection while substantially reducing parameters relative to a full $d_{\text{model}} \times d_{\text{model}}$ map.

With T overlapping blocks of length ℓ, pairwise attention costs $\mathcal{O}(T\,\ell^2 d_{\text{head}})$ and the windowed triadic pathway costs $\mathcal{O}(T\,\ell\,w^2 d_{\text{head}})$.

HOMA's block-and-window design induces a deliberate two-scale structure. Within each overlapping block, pairwise attention operates over the full $\ell \times \ell$ token-pair space; triadic attention is further restricted to a local window of size $w \leq \ell$. The pairwise pathway therefore operates at a more global scale capturing long-range dependencies, while the triadic pathway operates at a local scale capturing short-range higher-order cooperativities, and their combination aims to yield representations richer than either scale alone.

3 EXPERIMENTAL SETUP

All experiments use a shared pipeline in which only the attention mechanism varies. We compare PAIRWISE-2D, standard softmax self-attention; BLOCKWISE-2D, its overlapping-block variant; BLOCKWISE-3D, a purely triadic mechanism; and HOMA, our proposed pairwise–triadic attention mechanism. All experimental results are averaged across three seeds. Full hyperparameters and training setup are provided in Appendix A, Tables A.1–A.3. All experiments run on a single NVIDIA A100-SXM4-80GB GPU.

3.1 SYNTHETIC DIAGNOSTIC TASKS

Diagnostic models consist of an embedding layer, the attention mechanism under study, and a linear output classifier, with the standard post-attention feed-forward MLP omitted to limit additional nonlinear feature composition. Because HOMA includes an internal fusion MLP, we also report HOMA-ADD, which replaces learned fusion with a parameter-free sum of the pairwise and triadic outputs, isolating the contribution of learned fusion from that of combining the two interaction pathways. We consider two synthetic task families, PARITY–MAJORITY and MATCH2–MATCH3, to probe how interaction order and model capacity affect learnability under controlled conditions.

PARITY-k and MAJORITY-k. Given a random binary sequence $b \in \{0,1\}^L$, a set of k offsets $\mathcal{O} = \{o_1, \ldots, o_k\}$, and an interior position i, we define

$$y_i^{\text{PAR}} = \bigoplus_{j=1}^{k} b_{i+o_j}, \qquad y_i^{\text{MAJ}} = \mathbf{1}\left[\textstyle\sum_{j=1}^{k} b_{i+o_j} > k/2\right].$$

Here $\oplus$ denotes XOR. PARITY-k is an irreducible k-way interaction since every proper subset of its relevant inputs is statistically independent of the label. MAJORITY-k is a threshold on their additive sum and therefore serves as a lower-order control. The k offsets are spread evenly over the fixed interval $[-R, R]$, with $R = 3$, holding interaction reach fixed as k varies. Only positions whose full offset pattern lies within the sequence are labeled.

MATCH2 and MATCH3. For integer sequences $x \in \mathbb{Z}_M^N$ (Sanford et al., 2024),

$$y_i^{(2)} = \mathbf{1}[\exists j :\ x_i + x_j \equiv 0 \pmod{M}], \qquad y_i^{(3)} = \mathbf{1}[\exists j, z :\ x_i + x_j + x_z \equiv 0 \pmod{M}].$$

The modulus M is calibrated separately for each sequence length to maintain approximately balanced labels (Appendix D.1), and each integer is represented using a fixed Fourier basis. The modular-sum condition in MATCH3 admits a Fourier representation whose expansion contains products of three token-dependent coordinates, aligning naturally with the trilinear structure of triadic attention. We use MATCH3 as a controlled probe of explicit three-way interaction, with MATCH2 as the corresponding order-2 control.

For PARITY, we vary interaction order, embedding size, and depth; for MATCH3, we vary embedding size using a single attention layer. The single-layer setting is particularly informative because it removes cross-layer composition and more directly isolates differences between attention mechanisms. On both task families, block size equals sequence length, so PAIRWISE-2D and BLOCKWISE-2D are identical. The triadic window covers all task-relevant positions: $w = 2N - 1$ for MATCH2/MATCH3 and $w = 2R + 1 = 7$ for PARITY. Thus, receptive-field differences do not confound these comparisons; locality is examined separately in Section 5.

3.2 TAPE PROTEIN BENCHMARKS

We evaluate on three supervised tasks from the TAPE benchmark (Rao et al., 2019), spanning residue-level and pairwise prediction, and sequence-level regression. Models are trained without pretrained protein representations and use no multiple sequence alignments or coevolutionary features; the TAPE experiments are therefore designed as controlled comparisons of attention mechanisms rather than direct state-of-the-art comparisons.

Secondary Structure. Secondary Structure prediction is a residue-level classification task in which the model assigns each residue one of three structural labels: $\{\text{helix}, \text{strand}, \text{coil}\}$. Training and validation data are sourced from Klausen et al. (2019), and we evaluate on the held-out CB513 (Cuff & Barton, 1999), CASP12 (Moult et al., 2018), and TS115 (Yang et al., 2018) test sets. We report Q3 accuracy over non-padded residues.

Contact Prediction. Contact Prediction is a pairwise residue-classification task in which the model predicts a binary contact matrix $y \in \{0,1\}^{L \times L}$. Following TAPE (Rao et al., 2019), two residues are considered in contact if their Cα atoms lie within 8 Å in the experimentally resolved structure. Data are drawn from ProteinNet (AlQuraishi, 2019), with evaluation on the held-out CASP12 split (Moult et al., 2018). We report precision at $L/5$ over long-range residue pairs ($|i-j| \geq 24$).

Fluorescence. Fluorescence is a sequence-level regression task over mutated variants of green fluorescent protein (Sarkisyan et al., 2016), with log fluorescence intensity as the prediction target. Following TAPE (Rao et al., 2019), training variants lie within Hamming distance three of the parent sequence, whereas test variants contain at least four mutations. We report Spearman's rank correlation ρ.

4 RESULTS

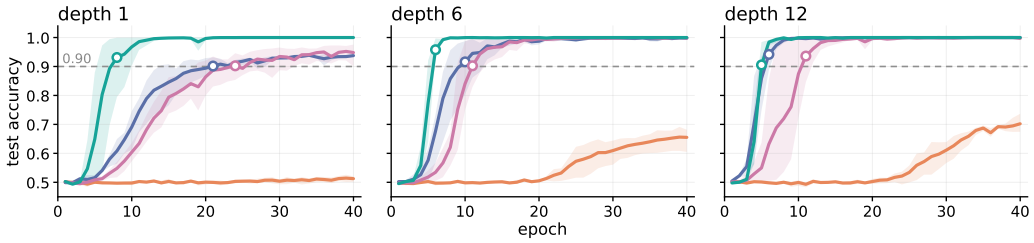
4.1 SYNTHETIC DIAGNOSTIC TASKS

Table 1 reports results on PARITY. All models solve the MAJORITY control and achieve perfect or near-perfect accuracy on PARITY tasks of orders 1–2 (Appendix C, Table C.1). PAIRWISE-2D can recover PARITY-3, but its accuracy degrades sharply on PARITY-4 and PARITY-5, and increasing width or depth only partially closes this gap within the training budget. BLOCKWISE-3D performs substantially better on the higher-order tasks and benefits from additional depth, while HOMA-ADD further improves performance in shallow and lower-dimensional regimes. HOMA with nonlinear fusion achieves perfect accuracy across all tested interaction orders, embedding dimensions, and depths.

HOMA's benefits on higher-order tasks extend beyond final accuracy to improved parameter efficiency and faster convergence. On PARITY-5 (Figure 2), HOMA solves the task already at depth 1, whereas PAIRWISE-2D fails within the training budget at every tested depth, and BLOCKWISE-3D reaches perfect accuracy at depth 12. Notably, with a single layer, HOMA-ADD still outperforms both baselines in final accuracy but requires about three times as many epochs as HOMA to reach 0.90 accuracy. The efficiency advantage is also evident on PARITY-3 and PARITY-4 with a single attention layer (Appendix C, Table C.2). Note that although both PAIRWISE-2D and HOMA achieve perfect accuracy on PARITY-3, HOMA requires approximately $5.4\times$ fewer parameters and converges in roughly half as many epochs. These results suggest that combining pairwise and triadic representations is advantageous for higher-order interaction learning, while their nonlinear fusion provides a particularly efficient mechanism for faster convergence when estimating dependencies beyond the primitive interaction order.

Table 1: **PARITY-k across interaction order k, network depth $D \in \{1, 6, 12\}$, and embedding size d.** Mean final test accuracy ± standard deviation over three seeds. Results for $k = 3$ are shown in gray since performance is near-saturated across attention mechanisms. HOMA results for $k = 4, 5$ are bold.

		k=3			k=4			k=5		
d	**mechanism**	D=1	6	12	D=1	6	12	D=1	6	12
32	PAIRWISE-2D	.978 ±.037	1.000 ±.000	1.000 ±.000	.499 ±.014	.729 ±.021	.694 ±.011	.503 ±.002	.611 ±.036	.601 ±.028
	BLOCKWISE-3D	1.000 ±.000	1.000 ±.000	1.000 ±.000	.964 ±.048	.997 ±.003	1.000 ±.000	.847 ±.125	.991 ±.011	1.000 ±.001
	HOMA-ADD	1.000 ±.000	1.000 ±.000	1.000 ±.000	.992 ±.010	1.000 ±.000	1.000 ±.000	.871 ±.066	.996 ±.007	.999 ±.002
	HOMA	1.000 ±.000	1.000 ±.000	1.000 ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000
64	PAIRWISE-2D	1.000 ±.000	1.000 ±.000	1.000 ±.000	.500 ±.004	.782 ±.101	.728 ±.039	.513 ±.010	.655 ±.043	.702 ±.032
	BLOCKWISE-3D	1.000 ±.000	1.000 ±.000	1.000 ±.000	.979 ±.018	1.000 ±.000	1.000 ±.000	.938 ±.001	.998 ±.003	1.000 ±.000
	HOMA-ADD	1.000 ±.000	1.000 ±.000	1.000 ±.000	.992 ±.011	1.000 ±.000	1.000 ±.000	.948 ±.041	.998 ±.003	.997 ±.005
	HOMA	1.000 ±.000	1.000 ±.000	1.000 ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000	**1.000** ±.000



	depth 1			depth 6			depth 12		
	params	acc	ep@0.90	params	acc	ep@0.90	params	acc	ep@0.90
Pairwise-2D	18k	0.513	>40	102k	0.655	>40	203k	0.702	>40
Blockwise-3D	19k	0.938	20.0	108k	0.998	9.7	215k	1.000	6.0
HOMA-add	19k	0.948	24.3	108k	0.998	11.3	215k	0.997	9.7
HOMA (fusion)	**25k**	**1.000**	**7.7**	146k	1.000	**5.7**	290k	1.000	**5.7**

Figure 2: **Convergence on PARITY-5 at embedding size $d = 64$ across network depths $D \in \{1, 6, 12\}$.** Curves show mean test accuracy over three seeds, with final values reported as `acc`; shaded bands span the minimum–maximum range across seeds. The dashed line marks 0.90 accuracy. Open circles and `ep@0.90` indicate the mean first-crossing epoch across seeds; > 40 indicates that no seed reached the threshold within 40 epochs.

Table 2: **MATCH3 across sequence lengths N and embedding sizes d_{model}.** Mean final test accuracy $\pm$ standard deviation over three seeds; M denotes the modulus. Bold marks the highest mean accuracy at each length and embedding size. $d_{0.90}$ is the smallest tested embedding size achieving mean accuracy ≥ 0.90. params reports the parameter count at $d_{0.90}$ relative to PAIRWISE-2D at the same embedding size. A dash means the threshold is never reached within the tested range, also indicated by $d_{0.90} > 256$. The two additional widths $d_{\text{model}} = 8$ and 256 are in Appendix D, Table D.2.

		test accuracy at d_{model}					
N (M)	mechanism	16	32	64	128	$d_{0.90}$	params
6 (30)	PAIRWISE-2D	0.816 $\pm$0.009	0.835 $\pm$0.001	0.830 $\pm$0.003	0.783 $\pm$0.009	>256	–
	BLOCKWISE-3D	0.836 $\pm$0.007	0.921 $\pm$0.012	0.958 $\pm$0.008	0.927 $\pm$0.028	32	1.08×
	HOMA-ADD	0.835 $\pm$0.003	0.929 $\pm$0.002	0.990 $\pm$0.002	0.975 $\pm$0.006	32	1.08×
	HOMA	**0.857** $\pm$0.006	**0.967** $\pm$0.007	**0.993** $\pm$0.002	**0.991** $\pm$0.003	32	1.60×
8 (56)	PAIRWISE-2D	0.708 $\pm$0.005	0.744 $\pm$0.004	0.761 $\pm$0.009	0.731 $\pm$0.012	>256	–
	BLOCKWISE-3D	0.726 $\pm$0.013	0.768 $\pm$0.005	**0.856** $\pm$0.004	0.912 $\pm$0.004	128	1.03×
	HOMA-ADD	0.734 $\pm$0.014	0.776 $\pm$0.004	0.848 $\pm$0.003	0.921 $\pm$0.004	128	1.03×
	HOMA	**0.745** $\pm$0.009	**0.799** $\pm$0.011	0.853 $\pm$0.003	**0.925** $\pm$0.008	128	1.19×

Table 2 reports results on MATCH3. Under the same protocol, MATCH2 achieves 1.000 accuracy over three seeds at every displayed width, except for one PAIRWISE-2D seed at $N = 8$, $d_{\text{model}} = 128$ (Appendix D, Table D.1). On MATCH3, increasing width yields relatively limited and non-monotonic gains for PAIRWISE-2D, whose best accuracy remains below 0.90. In contrast, all mechanisms with explicit triadic interactions cross 0.90 at $d = 32$ for $N = 6$ and $d = 128$ for $N = 8$. HOMA attains the highest overall accuracy in both settings, reaching 0.993 and 0.925 for $N = 6$ and $N = 8$, respectively.

4.2 TAPE PROTEIN BENCHMARKS

Figure 3a summarizes test performance for Secondary Structure (CB513), Contact Prediction, and Fluorescence. BLOCKWISE-2D tends to improve over PAIRWISE-2D across all three TAPE tasks. HOMA shows its clearest gains on Secondary Structure and Contact Prediction, with the gap over pairwise attention diminishing at the largest tested widths. On Secondary Structure, HOMA outperforms both pairwise and purely triadic baselines across all tested embedding sizes. On Contact

Prediction, BLOCKWISE-3D is stronger at smaller widths, while HOMA overtakes it at $d_{\text{model}} \geq 128$. Fluorescence shows substantially weaker separation, with BLOCKWISE-3D slightly outperforming HOMA overall. Full results across TAPE evaluations, including all Secondary Structure test sets, are provided in Appendix E, Table E.1.

Figure 3b compares the parameter count required to match HOMA's performance at each embedding size. On Secondary Structure, the strongest pairwise baseline requires up to $52\times$ more parameters to match HOMA, while BLOCKWISE-3D requires up to $15\times$ more parameters. On Contact Prediction, the advantage is primarily over pairwise attention, which requires up to $28\times$ more parameters. Full performance results across embedding widths are reported in Appendix E. HOMA incurs greater memory cost at a fixed embedding size because of its triadic pathway, but this overhead can be offset in some matched-performance settings: on Secondary Structure, HOMA at $d_{\text{model}} = 64$ uses approximately 2.97 GB, whereas matching it requires BLOCKWISE-2D at $d_{\text{model}} = 512$ (3.55 GB) and BLOCKWISE-3D at $d_{\text{model}} = 128$ (3.11 GB). Full throughput and peak-memory results at every width are provided in Appendix F, Table F.1.

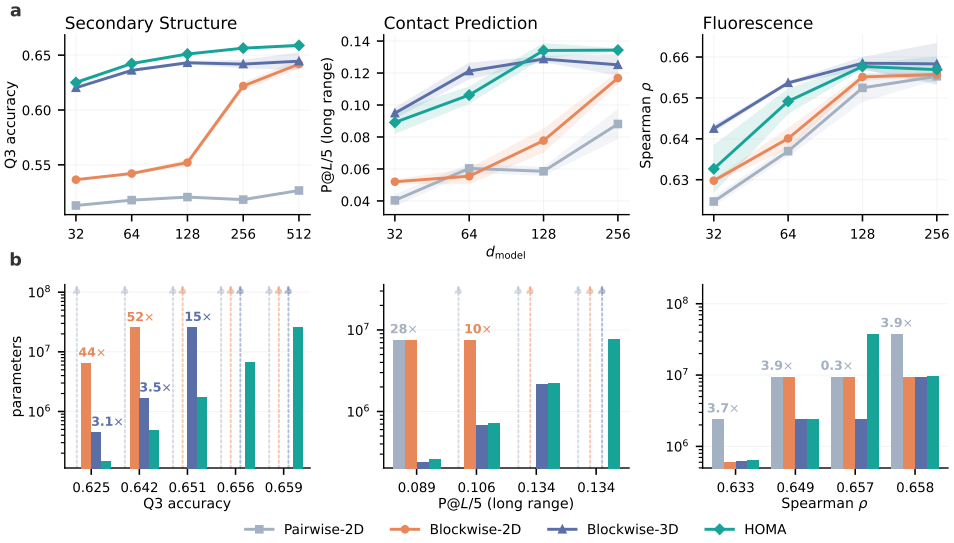


Figure 3: **Test performance and performance-matched parameter efficiency across TAPE Secondary Structure (CB513), Contact Prediction, and Fluorescence.** Results are averaged over three seeds. **(a)** Mean performance across embedding sizes, with shaded regions showing ± 1 standard deviation. **(b)** For each attention mechanism, we report the minimum parameter count across tested widths that matches each HOMA performance target. A run is considered to match a target if its score is within 1.5 pooled seed standard deviations of the HOMA score or within 0.5% in relative terms. Dotted arrows denote targets not reached; annotations show parameter count relative to HOMA. The parameter axis is logarithmic.

5 ABLATIONS AND ANALYSIS

We conduct ablations on PARITY-3 to examine how window size and depth affect the receptive field of the triadic pathway. We define the interaction reach R as the maximum absolute offset from the query position among the positions entering the PARITY-3 interaction, and the triadic window reach as $R_w = (w-1)/2$, so a window of size w covers offsets up to $\pm R_w$. Accordingly, Figure 4(a,b) shows that BLOCKWISE-3D achieves perfect accuracy when $R \leq R_w$ but drops to chance beyond this range. HOMA avoids this collapse because its pairwise branch remains available, maintaining or exceeding the PAIRWISE-2D baseline. Depth partially mitigates this limitation, as shown in Figure 4(c): $R = 2$ is solved at depth 2, while $R = 3$ only partially recovers at depth 4. Overall, the triadic pathway degrades when the interaction reach exceeds its local window reach, although additional depth can recover some longer-range dependencies. The corresponding numerical results are provided in Appendix G, Tables G.1–G.2.

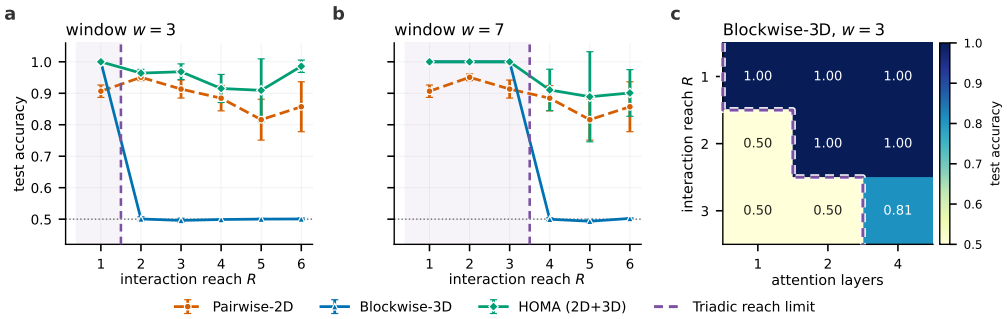


Figure 4: **Triadic window coverage on PARITY-3.** All panels use dedicated coverage-sweep runs with $d_{\text{model}} = 32$ and sequence length $L = 24$; results are averaged over three seeds. **(a,b)** Test accuracy across interaction reach R for window sizes w=3 and w=7; the shaded region and dashed line indicate the triadic reach limit R_w. **(c)** BLOCKWISE-3D accuracy across interaction reach and depth for w=3.

Table 3: **Component teardown of HOMA on PARITY-k and MATCH3.** Each row ablates one component with the others fixed. Results use one attention layer, d_{model}=32, and three seeds. acc is mean final test accuracy; ep@0.90 is the mean epoch to reach 0.90 accuracy among successful seeds within 40 epochs. † and ‡ denote one and two successful seeds, respectively; > 40 denotes no successful seed.

	Component			PARITY-3		PARITY-4		PARITY-5		MATCH3	
Design	pooling	third factor	fusion	acc	ep@0.90	acc	ep@0.90	acc	ep@0.90	acc	ep@0.90
HOMA	**softmax**	$W^{(U)}$	**MLP**	**1.000**	**5.7**	**1.000**	**7.7**	**1.000**	**9.3**	**0.967**	**10.7**
Replace	uniform	$W^{(U)}$	MLP	1.000	9.7	0.994	16.3	0.955	21.7	0.861	>40
	softmax	$W^{(U)}$=$W^{(K)}$	MLP	1.000	4.3	1.000	7.3	0.983	9.7	0.933	19.3
Replace	softmax	$W^{(U)}$	sum	1.000	7.3	0.992	17.0	0.871	32.0†	0.929	19.3
	uniform	$W^{(U)}$	sum	0.988	18.7	0.495	>40	0.500	>40	0.838	>40
	softmax	$W^{(U)}$=$W^{(K)}$	sum	1.000	6.3	0.978	23.3	0.909	25.0‡	0.904	32.0

Table 3 isolates three design choices in HOMA. We remove selective triadic attention by replacing it with uniform pooling, remove the independent parameterization of the third projection by equating $W^{(U)}$ with $W^{(K)}$, and remove the learned fusion MLP by replacing it with parameter-free addition. Removing the fusion layer substantially slows convergence, especially on PARITY-4 and PARITY-5, with accuracy also dropping. With the fusion MLP retained, uniform pooling mainly affects convergence speed, with accuracy also decreasing on MATCH3. The distinction becomes clearer without the fusion layer, where uniform pooling reduces PARITY-4 and PARITY-5 to chance, matching the pairwise baseline, while also substantially slowing convergence. Tying U to K has a smaller and less consistent effect across settings. Overall, the fusion layer is beneficial for faster convergence, while selective triadic attention provides the clearest contribution to HOMA on higher-order tasks.

6 CONCLUSION

We present HOMA, a Higher-Order Modular Attention mechanism that combines pairwise and triadic pathways with localized higher-order computation to make triadic attention tractable. Across controlled experiments, HOMA becomes increasingly effective as interaction order grows, with learned nonlinear fusion further improving convergence. On TAPE, these advantages extend to natural sequence data in a task-dependent manner. Overall, the results support HOMA as a useful inductive bias when task structure extends beyond pairwise interactions. Future work will address the computational and locality constraints of the triadic pathway and evaluate HOMA in pretrained models and broader sequence domains.

AI USE STATEMENT

Generative AI tools were used for writing assistance, including copy-editing, rephrasing, and LaTeX formatting of text drafted by the authors, and to assist with code implementation and debugging. All AI-assisted text and code were reviewed, tested, and verified by the authors. The authors take full responsibility for the final content of this work, including the implementation, reported results, text, claims, and released artifacts.

ETHICS STATEMENT

This work uses only publicly released protein benchmark datasets distributed through TAPE (Rao et al., 2019) and involves no human subjects, personally identifying information, or sensitive data. The released models are small-scale attention-mechanism prototypes trained on public benchmarks and pose no foreseeable risk of misuse. We are aware of no conflicts of interest associated with this work.

REPRODUCIBILITY STATEMENT

All architectural, optimizer, and attention-specific hyperparameters are reported in Section 3 and in full in Appendix A, Tables A.1–A.3. Section 3 specifies the dataset splits, tokenization, evaluation metrics, and profiling methodology for all three tasks. All results are reported as mean $\pm$ standard deviation over three fixed random seeds, run on a single NVIDIA A100-SXM4-80GB GPU. An anonymized code package containing model implementations and training and evaluation scripts, together with instructions for reproducing the reported experiments, is described in Appendix H.

REFERENCES

Mohammed AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311, 2019. doi: 10.1186/s12859-019-2932-0.

Song Bai, Feihu Zhang, and Philip H. S. Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.576.

Sayak Chakrabarti, Toniann Pitassi, and Josh Alman. Poly-attention: A general scheme for higher-order self-attention. *arXiv preprint arXiv:2602.02422*, 2026.

Avijit Chakrabartty, Andrew J. Doig, and Robert L. Baldwin. Helix capping propensities in peptides parallel those in proteins. *Proceedings of the National Academy of Sciences*, 90(23):11332–11336, 1993. doi: 10.1073/pnas.90.23.11332.

David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.527.

James Clift and Daniel Murfet. Logic and the 2-simplicial transformer. In *International Conference on Learning Representations*, 2020.

James A. Cuff and Geoffrey J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 34(4): 508–519, 1999. doi: 10.1002/(SICI)1097-0134(19990301)34:4<508::AID-PROT10>3.0.CO;2-4.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020. doi: 10.1162/tacl_a_00306.

Amnon Horovitz and Alan R. Fersht. Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *Journal of Molecular Biology*, 214(3):613–617, 1990.

IUPAC-IUB Joint Commission on Biochemical Nomenclature. Nomenclature and symbolism for amino acids and peptides (recommendations 1983). *Pure and Applied Chemistry*, 56(5):595–624, 1984. doi: 10.1351/pac198456050595.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, et al. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. *arXiv preprint arXiv:2407.02490*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Soenderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 2019.

Alexander Kozachinskiy, Felipe Urrutia, Hector Jimenez, Tomasz Steifer, Germán Pizarro, Matías Fuentes, Francisco Meza, Cristian B. Calderon, and Cristóbal Rojas. Strassen attention, split VC dimension and compositionality in transformers. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2501.19215.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011. doi: 10.1073/pnas.1111471108.

John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018. ISSN 08873585. doi: 10.1002/prot.25415. URL `http://doi.wiley.com/10.1002/prot.25415`.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture. In *Proceedings of the 1st Conference on Language Modeling*, 2024.

Frank J. Poelwijk, Vijay Krishna, and Rama Ranganathan. The context-dependence of mutations: A linkage of formalisms. *PLOS Computational Biology*, 12(6):e1004771, 2016. doi: 10.1371/journal.pcbi.1004771.

Frank J. Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature Communications*, 10:4213, 2019. doi: 10.1038/s41467-019-12130-8.

Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. In *Findings of the Association for Computational Linguistics*, 2020. URL `https://arxiv.org/abs/1911.02972`.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, 2019.

Zachary R. Sailer and Michael J. Harms. High-order epistasis shapes evolutionary trajectories. *PLOS Computational Biology*, 13(5):e1005541, 2017. doi: 10.1371/journal.pcbi.1005541.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems*, 2024.

Karen S. Sarkisyan, Dmitry A. Bolotin, Margarita V. Meer, Dinara R. Usmanova, Alexander S. Mishin, George V. Sharonov, Dmitry N. Ivankov, Nina G. Bozhanova, Mikhail S. Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397, 2016.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Bi-directional block self-attention for fast and memory-efficient sequence modeling. In *International Conference on Learning Representations (ICLR)*, 2018. URL `https://arxiv.org/abs/1804.00857`.

James A. Smith and Linda G. Pease. Reverse turns in peptides and proteins. *CRC Critical Reviews in Biochemistry*, 8(4):315–399, 1980. doi: 10.3109/10409238009105479.

The UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019. doi: 10.1093/nar/gky1049.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

C. M. Venkatachalam. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, 6(10):1425–1436, 1968. doi: 10.1002/bip.1968.360061006.

Daniel M. Weinreich, Y. Lan, Christopher S. Wylie, and Robert B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, 23 (6):700–707, 2013.

James A. Wells. Additivity of mutational effects in proteins. *Biochemistry*, 29(37):8509–8517, 1990.

Nicholas C. Wu, Liang Dai, Catherine A. Olson, James O. Lloyd-Smith, and Ren Sun. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5:e16965, 2016. doi: 10.7554/eLife.16965.

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics*, 19(3):482–494, 2018. doi: 10.1093/bib/bbw129.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, 2020.

APPENDIX

We provide experimental details, full results, and reproducibility information omitted from the main paper. The appendices are organized as follows:

- Appendix A: complete hyperparameter specification.
- Appendix B: diagnostic-task protocol and conventions.
- Appendix C: additional PARITY results.
- Appendix D: MATCH2 and MATCH3 setup and results.
- Appendix E: full TAPE capacity results.
- Appendix F: compute and peak-memory measurements.
- Appendix G: triadic-window coverage and depth.
- Appendix H: code package and reproducibility details.

A COMPLETE HYPERPARAMETER SPECIFICATION

Tables A.1–A.3 give the complete experimental configuration. Within each task's capacity sweep, only d_{model} varies.

Table A.1: **Data and task specification.** "—" denotes a non-applicable setting. The MATCH modulus M is calibrated by sequence length and task order (Section D.1).

	PARITY/ MAJORITY	MATCH2/ MATCH3	Secondary Structure	Contact Prediction	Fluorescence
Train / valid / test	3,000 / — / 800	30,000 / — / 2,000	8,678 / 2,170 / 3 sets	25,299 / 224 / 40	21,446 / 5,362 / 27,217
Sequence length	L=16	$N \in \{6, 8\}$	pad to 512	crop to 256	fixed at 237
Vocabulary	2 (binary)	M (calibrated)	30 (IUPAC)	30 (IUPAC)	30 (IUPAC)
Label	per position	per position	per residue	per pair	per sequence
Test split	held out	held out	CB513 / CASP12 / TS115	CASP12	$H_d \geq 4$

Table A.2: **Architecture and attention.** Only the attention mechanism varies within a task; every other quantity here is shared across mechanisms at a given width.

	PARITY/ MAJORITY	MATCH2/ MATCH3	Secondary Structure	Contact Prediction	Fluorescence
d_{model} range	8–128 1, 6, 12 (grid); 1 (capacity);	8–256	32–512	32–256	32–256
Layers	1, 2, 4 (coverage)	1	12	12	12
Heads	4	4	8	8	8
FFN width	none	none	$2d_{\text{model}}$	$2d_{\text{model}}$	$2d_{\text{model}}$
Dropout	0	0	0.4	0.1	0.1
Task head	linear, 2-way	linear, 2-way	per-residue cls.	pairwise, symmetric	mean-pool reg.
Block ℓ / stride s	single block	ℓ=s=N	30 / 15	32 / 16	30 / 15
Triadic window w	7 (uniform across k)	$2N-1$	5	5	5
Low-rank U rank r	8	8	8	8	8

Losses, stated per task. Ignore index -100 masks unlabelled or padded token positions, and -1 masks unscored residue pairs in Contact Prediction. Contact Prediction upweights positives by $5\times$ because contacts comprise approximately 2% of scored pairs.

Checkpoint selection on the TAPE tasks. TAPE results use the task-specific validation criterion in Table A.3, except Contact Prediction, which reports the final epoch; test data are never used for selection. Diagnostic results instead use the final epoch of a fixed budget with no checkpoint selection (Section B).

Table A.3: **Optimization, loss, and evaluation settings.**

	PARITY/ MAJORITY	MATCH2/ MATCH3	Secondary Structure	Contact Prediction	Fluorescence
Optimizer	Adam	Adam	Adam	AdamW	Adam
Learning rate	2×10^{-3}	3×10^{-3}	10^{-4}	10^{-4}	10^{-4}
Weight decay	0	0	0	0.01	0
LR schedule	none	none	none	OneCycle (10% warm-up)	cosine (6% warm-up)
Gradient clipping	none	none	none	1.0	1.0
Batch size	128	256	16	8	16
Epochs	40	40	10	8	10
Loss	CE, ignore -100	CE	token CE, ignore -100	weighted CE, ignore -1	MSE
Class weighting	—	—	—	$5\times$ positive	—
Metric	accuracy	accuracy	Q3 accuracy	P@$L/5$ long-range	Spearman ρ
Checkpoint	last epoch	last epoch	best val. loss	last epoch	best val. ρ
Seeds	0, 1, 2	0, 1, 2	42, 456, 608	0, 1, 2	42, 456, 608

Notes on individual settings. For the diagnostic sweeps, PAIRWISE-2D and BLOCKWISE-2D are identical because the length-16 sequence forms one block; their per-seed accuracies agree, so the tables show one pairwise row. A single triadic window, determined by the widest offset pattern, is held fixed across k; Section G varies this window separately.

For Contact Prediction, cropping is a random window during training and the centred window at evaluation, so validation and test are deterministic. Proteins shorter than 32 residues are dropped, as they cannot hold a scoreable $|i - j| \geq 6$ pair.

The MATCH tasks. The frozen Fourier embedding is followed by a learned linear projection, and positional embeddings are disabled. Runs use independent train and test streams for each seed, no validation split, and the final epoch of the 40-epoch budget. Tested widths are 16–128 for MATCH2 and 8–256 for MATCH3; majority-class accuracy is 0.500–0.540.

Preprocessing and sequence handling. All protein datasets use residue-level tokenization with the TAPE IUPAC tokenizer, a 30-token vocabulary of five control symbols and a 25-character residue alphabet: the 20 standard amino acids, selenocysteine and pyrrolysine, two ambiguity codes, and one unknown-residue symbol (IUPAC-IUB Joint Commission on Biochemical Nomenclature, 1984; The UniProt Consortium, 2019). Secondary Structure sequences are padded to 512, Contact Prediction sequences are cropped to 256, and Fluorescence inputs are fixed at 237 tokens. Padded positions are excluded from attention and evaluation by masking.

A note on the epoch budget. Epochs are equal in number but not in optimization steps across the TAPE tasks. At batch size 16, one epoch is 543 steps for Secondary Structure and 1,341 for Fluorescence. Comparisons within a task are unaffected; comparisons of convergence across tasks are not meaningful and should not be drawn.

B PROTOCOL AND CONVENTIONS FOR THE DIAGNOSTIC TASKS

Diagnostic results use last-epoch accuracy over labelled test positions; chance is 0.50. Selecting the best epoch instead moves no cell by more than 0.010; the two cells that move by more than 0.005, PAIRWISE-2D at $k = 4$ and 5, are at chance under either convention.

Epochs-to-threshold is the first test epoch above the specified accuracy. Means include only successful seeds and show the success count, so values based on fewer than three seeds are conditional on success. "$>$40" means that no seed crossed within the training budget.

Two conventions that must not be mixed. Tables report the mean crossing epoch among successful seeds, whereas Figure 2 assigns non-crossing seeds epoch 41 and averages all seeds. Consequently, figure annotations and table entries for the same configuration may differ.

C PARITY RESULTS

Table C.1: **Lower-order PARITY, the MAJORITY control, and convergence.** **(a)** Mean final test accuracy ± standard deviation for PARITY-k, $k \in \{1, 2\}$. **(b)** Mean final accuracy for MAJORITY-k; all standard deviations are below 0.001. **(c)** Mean epochs to 0.90 among successful seeds for PARITY-k at d_{model}=64; counts are shown when fewer than three seeds cross, ">40" denotes no crossing, and *n/a* denotes an unavailable learning curve (Section B).

(a) PARITY-k, k=1, 2

		k=1			k=2		
d	mechanism	D=1	6	12	D=1	6	12
32	PAIRWISE-2D	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000
	BLOCKWISE-3D	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000
	HOMA-ADD	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000
	HOMA	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000
64	PAIRWISE-2D	1.000 ±.000	1.000 ±.000	1.000 ±.000	0.983 ±.030	1.000 ±.000	1.000 ±.000
	BLOCKWISE-3D	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000
	HOMA-ADD	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000
	HOMA	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000	1.000 ±.000

(b) MAJORITY-k control across network depth D and embedding size d

		k=1			k=2			k=3			k=4			k=5		
d	mechanism	D=1	6	12	D=1	6	12	D=1	6	12	D=1	6	12	D=1	6	12
32	PAIRWISE-2D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BLOCKWISE-3D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	HOMA-ADD	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	HOMA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
64	PAIRWISE-2D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	BLOCKWISE-3D	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	HOMA-ADD	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	HOMA	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

(c) Epochs to 0.90 on PARITY-k, d_{model}=64

	k=3			k=4			k=5		
mechanism	D=1	6	12	D=1	6	12	D=1	6	12
PAIRWISE-2D	*n/a*	10.3	10.0	>40	>40	>40	>40	>40	>40
BLOCKWISE-3D	*n/a*	3.0	2.7	12.3	6.7	6.0	20.0	9.7	6.0
HOMA-ADD	4.3	4.0	5.0	12.7	10.3	10.7	24.3	11.3	9.7
HOMA	*n/a*	3.0	3.0	5.3	4.3	4.0	7.7	5.7	5.7

Table C.2: **Capacity on PARITY-3 and PARITY-4.** Trainable parameters, mean final test accuracy over three seeds, and mean epochs to 0.90 among successful seeds. One attention layer without a residual connection; a separate run series from the depth grid of Table C.1. The order-4 ladder is a separate run of the same sweep under this protocol; at d_{model}=32, where it overlaps the original, the two agree to within 0.008 accuracy.

	PAIRWISE-2D			BLOCKWISE-3D			HOMA		
d_{model}	params	acc	ep@0.90	params	acc	ep@0.90	params	acc	ep@0.90
PARITY-3									
8	466	0.586 $\pm$.041	>40	594	0.946 $\pm$.058	22.5 (2/3)	1,492	0.982 $\pm$.030	19.3
16	1,442	0.879 $\pm$.082	35.5 (2/3)	1,698	0.989 $\pm$.019	10.0	3,366	1.000 $\pm$.000	6.0
32	4,930	0.978 $\pm$.037	21.0	5,442	1.000 $\pm$.000	4.7	8,650	1.000 $\pm$.000	5.0
64	18,050	1.000 $\pm$.000	11.7	19,074	1.000 $\pm$.000	3.0	25,362	1.000 $\pm$.000	3.0
PARITY-4									
8	466	0.501 $\pm$.006	>40	594	0.785 $\pm$.172	23.0 (1/3)	1,492	1.000 $\pm$.001	19.7
16	1,442	0.499 $\pm$.007	>40	1,698	0.821 $\pm$.036	>40	3,366	1.000 $\pm$.000	9.0
32	4,930	0.512 $\pm$.004	>40	5,442	0.929 $\pm$.062	22.5 (2/3)	8,650	1.000 $\pm$.000	5.7
64	18,050	0.513 $\pm$.014	>40	19,074	0.983 $\pm$.026	9.7	25,362	1.000 $\pm$.000	3.7

D MATCH2 AND MATCH3: SETUP AND RESULTS

D.1 TASK SETUP, AND WHERE M AND N STOP

Task. The task definitions are given in Section 3. Here j and z range over the full sequence and may equal each other or i. Each token is encoded by a frozen sine–cosine Fourier feature followed by a learned projection; d_{model} therefore limits the represented frequencies.

Choice of the modulus M. For approximately uniform tokens, $\Pr(y_i{=}0) \approx e^{-L_N/M}$, so balanced labels require $M \approx L_N/\ln 2$, where L_N is the number of candidate partners per query: $L_N = N$ for MATCH2 and $L_N = \binom{N+1}{2}$ for MATCH3. Hence $M = \Theta(N)$ for MATCH2 and $M = \Theta(N^2)$ for MATCH3. For each (N, order), we select the grid value of M whose measured positive rate is closest to 0.5; all reported rates lie within 0.04 of 0.50.

Why the reported lengths stop at N=8. A complete real Fourier basis requires approximately $M-1$ coordinates. Because MATCH3 uses $M = \Theta(N^2)$, its required embedding width also grows quadratically with N. We therefore stop at $N = 8$: the tested widths include a complete basis for $N = 6$ ($M = 30$) and $N = 8$ ($M = 56$), whereas longer sequences would confound attention order with incomplete input features.

D.2 RESULTS

Table D.1: **MATCH2 control.** Mean final test accuracy $\pm$ standard deviation over three seeds across sequence lengths N, calibrated moduli M, and embedding widths d_{model}.

		test accuracy at d_{model}			
N (M)	mechanism	16	32	64	128
6 (9)	PAIRWISE-2D	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000
	BLOCKWISE-3D	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000
	HOMA-ADD	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000
	HOMA	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000
8 (14)	PAIRWISE-2D	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	0.979 $\pm$.036
	BLOCKWISE-3D	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000
	HOMA-ADD	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000
	HOMA	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000

Table D.2: **Extended MATCH3 width sweep.** Mean final test accuracy ± standard deviation over three seeds. Widths 16–128 reproduce Table 2 for completeness; widths 8 and 256 are additional.

		test accuracy at d_{model}					
N (M)	mechanism	8	16	32	64	128	256
6 (30)	PAIRWISE-2D	0.760 ±.008	0.816 ±.009	0.835 ±.001	0.830 ±.003	0.783 ±.009	0.776 ±.026
	BLOCKWISE-3D	0.770 ±.015	0.836 ±.007	0.921 ±.012	0.958 ±.008	0.927 ±.028	0.774 ±.008
	HOMA-ADD	0.741 ±.038	0.835 ±.003	0.929 ±.002	0.990 ±.002	0.975 ±.006	0.817 ±.010
	HOMA	0.786 ±.009	0.857 ±.006	0.967 ±.007	0.993 ±.002	0.991 ±.003	0.819 ±.015
8 (56)	PAIRWISE-2D	0.625 ±.014	0.708 ±.005	0.744 ±.004	0.761 ±.009	0.731 ±.012	0.726 ±.007
	BLOCKWISE-3D	0.618 ±.012	0.726 ±.013	0.768 ±.005	0.856 ±.004	0.912 ±.004	0.717 ±.018
	HOMA-ADD	0.614 ±.015	0.734 ±.014	0.776 ±.004	0.848 ±.003	0.921 ±.004	0.846 ±.089
	HOMA	0.629 ±.022	0.745 ±.009	0.799 ±.011	0.853 ±.003	0.925 ±.008	0.722 ±.019

E TAPE CAPACITY STUDY: FULL RESULTS

Table E.1: **TAPE results by task and width.** Mean ± standard deviation over three seeds; the best mechanism at each width is bold. Secondary Structure reports Q3 accuracy on CB513, CASP12, and TS115; Figure 3 plots CB513. Contact Prediction reports long-range P@$L/5$, and Fluorescence reports Spearman ρ. "–" denotes an untested width.

		score at d_{model}				
task	mechanism	32	64	128	256	512
Secondary Structure (CB513)	PAIRWISE-2D	0.513 ±.001	0.518 ±.001	0.521 ±.001	0.518 ±.002	0.527 ±.001
	BLOCKWISE-2D	0.536 ±.001	0.542 ±.002	0.552 ±.004	0.622 ±.004	0.642 ±.005
	BLOCKWISE-3D	0.620 ±.001	0.636 ±.002	0.643 ±.002	0.642 ±.005	0.644 ±.010
	HOMA	**0.625** ±.003	**0.642** ±.001	**0.651** ±.001	**0.656** ±.001	**0.659** ±.003
Secondary Structure (CASP12)	PAIRWISE-2D	0.542 ±.010	0.543 ±.005	0.539 ±.003	0.544 ±.006	0.538 ±.011
	BLOCKWISE-2D	0.570 ±.006	0.586 ±.005	0.588 ±.004	0.639 ±.006	0.658 ±.003
	BLOCKWISE-3D	0.625 ±.007	0.635 ±.006	0.648 ±.006	0.653 ±.006	0.651 ±.009
	HOMA	**0.635** ±.007	**0.650** ±.004	**0.665** ±.001	**0.664** ±.010	**0.669** ±.007
Secondary Structure (TS115)	PAIRWISE-2D	0.557 ±.001	0.564 ±.001	0.563 ±.002	0.563 ±.002	0.569 ±.002
	BLOCKWISE-2D	0.574 ±.003	0.582 ±.001	0.588 ±.005	0.655 ±.002	0.672 ±.009
	BLOCKWISE-3D	0.646 ±.005	0.659 ±.003	0.672 ±.003	0.671 ±.006	0.675 ±.008
	HOMA	**0.653** ±.004	**0.669** ±.002	**0.679** ±.005	**0.685** ±.002	**0.691** ±.002
Contact Prediction (P@$L/5$)	PAIRWISE-2D	0.040 ±.005	0.060 ±.004	0.059 ±.003	0.088 ±.012	–
	BLOCKWISE-2D	0.052 ±.002	0.055 ±.006	0.078 ±.009	0.117 ±.005	–
	BLOCKWISE-3D	**0.095** ±.004	**0.121** ±.006	0.129 ±.002	0.125 ±.008	–
	HOMA	0.089 ±.009	0.106 ±.007	**0.134** ±.006	**0.134** ±.001	–
Fluorescence (ρ)	PAIRWISE-2D	0.625 ±.001	0.637 ±.002	0.652 ±.004	0.655 ±.001	–
	BLOCKWISE-2D	0.630 ±.001	0.640 ±.003	0.655 ±.001	0.656 ±.001	–
	BLOCKWISE-3D	**0.642** ±.001	**0.654** ±.001	**0.658** ±.002	**0.658** ±.006	–
	HOMA	0.633 ±.007	0.649 ±.004	0.658 ±.001	0.657 ±.004	–

F COMPUTE AND PEAK MEMORY

Table F.1 reports trainable parameters, throughput, and peak allocated GPU memory for all tested configurations.

The overhead at fixed width is real. At d_{model}=512 on Secondary Structure, HOMA processes 31.5k residues/s and uses 14.80 GB, compared with 75.9k residues/s and 3.55 GB for BLOCKWISE-2D. Parameter counts differ by less than 2% (25.9M versus 25.5M), so the principal fixed-width overhead is compute and memory.

Table F.1: **Parameters, throughput, and peak memory by task and width.** *par*: trainable parameters; *tput*: thousands of residues per second; *mem*: peak allocated GB. Values are means over three seeds. Throughput is compute-only for Secondary Structure and Fluorescence and wall-clock-derived for Contact Prediction, so comparisons are valid within, but not across, tasks. "–" denotes an untested width.

	PAIRWISE-2D			BLOCKWISE-2D			BLOCKWISE-3D			HOMA		
d_{model}	par	tput	mem	par	tput	mem	par	tput	mem	par	tput	mem
Secondary Structure												
32	120k	165	2.04	120k	217	0.42	126k	138	0.91	146k	102	2.14
64	437k	159	2.19	437k	215	0.60	449k	129	1.64	487k	97.1	2.97
128	1.7M	147	2.50	1.7M	216	0.99	1.7M	112	3.11	1.8M	84.3	4.63
256	6.5M	101	3.14	6.5M	137	1.81	6.5M	72.0	6.08	6.7M	57.1	8.02
512	25.5M	68.2	4.53	25.5M	75.9	3.55	25.6M	37.8	12.05	25.9M	31.5	14.80
Contact Prediction												
32	235k	64.5	0.56	235k	59.9	0.42	241k	46.7	0.56	261k	40.0	0.88
64	666k	65.6	0.87	666k	57.5	0.75	679k	46.5	1.04	717k	39.2	1.40
128	2.1M	70.9	1.52	2.1M	63.4	1.45	2.1M	47.0	2.01	2.2M	36.6	2.45
256	7.4M	53.3	2.84	7.4M	51.3	2.79	7.4M	36.9	3.95	7.6M	32.8	4.48
Fluorescence												
32	612k	101	0.91	612k	79.4	0.60	618k	54.1	0.81	638k	42.6	1.34
64	2.4M	96.6	1.00	2.4M	76.7	0.71	2.4M	53.0	1.14	2.4M	42.4	1.73
128	9.5M	96.4	1.24	9.5M	77.5	0.99	9.5M	52.9	1.88	9.6M	42.1	2.55
256	37.6M	95.9	1.91	37.6M	76.6	1.46	37.6M	52.1	3.11	37.8M	41.3	3.94

G TRIADIC WINDOW COVERAGE AND DEPTH

Tables G.1 and G.2 provide the numerical results underlying Figure 4.

Table G.1: **Triadic-window coverage on PARITY-3 at depth 1.** Dedicated coverage-sweep runs use $d_{\text{model}} = 32$ and sequence length $L = 24$. The window reach is $R_w = (w-1)/2$. Mean final test accuracy $\pm$ standard deviation over three seeds. PAIRWISE-2D is independent of w, so the same runs are shown for both window settings.

		interaction reach R					
window	mechanism	1	2	3	4	5	6
w=3 (R_w=1)	PAIRWISE-2D	0.907 $\pm$.024	0.951 $\pm$.013	0.913 $\pm$.035	0.884 $\pm$.049	0.816 $\pm$.080	0.857 $\pm$.097
	BLOCKWISE-3D	1.000 $\pm$.000	0.501 $\pm$.009	0.496 $\pm$.008	0.499 $\pm$.001	0.500 $\pm$.003	0.501 $\pm$.004
	HOMA-ADD	1.000 $\pm$.000	0.858 $\pm$.065	0.936 $\pm$.071	0.862 $\pm$.074	0.854 $\pm$.046	0.829 $\pm$.065
	HOMA	1.000 $\pm$.000	0.964 $\pm$.015	0.968 $\pm$.031	0.915 $\pm$.055	0.909 $\pm$.122	0.986 $\pm$.024
w=7 (R_w=3)	PAIRWISE-2D	0.907 $\pm$.024	0.951 $\pm$.013	0.913 $\pm$.035	0.884 $\pm$.049	0.816 $\pm$.080	0.857 $\pm$.097
	BLOCKWISE-3D	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	0.500 $\pm$.006	0.493 $\pm$.008	0.502 $\pm$.002
	HOMA-ADD	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	0.942 $\pm$.065	0.884 $\pm$.064	0.846 $\pm$.087
	HOMA	1.000 $\pm$.000	1.000 $\pm$.000	1.000 $\pm$.000	0.911 $\pm$.081	0.889 $\pm$.175	0.901 $\pm$.091

Table G.2: **Accuracy by reach and depth for $w = 3$ ($R_w = 1$).** Dedicated coverage-sweep runs use $d_{\text{model}} = 32$ and sequence length $L = 24$. Mean final test accuracy $\pm$ standard deviation over three seeds.

		attention layers		
reach	mechanism	1	2	4
R=1	BLOCKWISE-3D	1.000 $\pm.000$	1.000 $\pm.000$	1.000 $\pm.000$
	HOMA-ADD	1.000 $\pm.000$	1.000 $\pm.000$	1.000 $\pm.000$
	HOMA	1.000 $\pm.000$	1.000 $\pm.000$	1.000 $\pm.000$
R=2	BLOCKWISE-3D	0.501 $\pm.009$	1.000 $\pm.000$	1.000 $\pm.000$
	HOMA-ADD	0.858 $\pm.065$	1.000 $\pm.000$	1.000 $\pm.000$
	HOMA	0.964 $\pm.015$	0.972 $\pm.045$	1.000 $\pm.000$
R=3	BLOCKWISE-3D	0.496 $\pm.008$	0.499 $\pm.004$	0.811 $\pm.065$
	HOMA-ADD	0.936 $\pm.071$	0.964 $\pm.040$	0.975 $\pm.044$
	HOMA	0.968 $\pm.031$	0.903 $\pm.046$	1.000 $\pm.000$

H CODE PACKAGE AND REPRODUCIBILITY

The anonymized package contains the model implementations, the diagnostic-task generators, the TAPE data pipelines, and one training script per set of runs, with its protocol fixed in the script. Each script prints its tables from the results it writes.

Seeds. Standard deviations use the sample $(n - 1)$ definition. Before aggregation, 72 duplicated PARITY records at $k = 3$ and $d_{\text{model}} \in \{32, 64\}$ are removed.