Higher-Order Modular Attention: Task-Aware Transformer Architectures for Protein Sequence Modeling

Shirin Amiraslani
Department of Mathematics and Statistics
York University
Toronto, ON M3J 1P3, Canada
shirinamiraslani@gmail.com

Xin Gao
Department of Mathematics and Statistics
York University
Toronto, ON M3J 1P3, Canada
xingao@yorku.ca

Abstract

Since the Transformer, many attention variants have improved efficiency and context modeling, yet most still rely on pairwise token interactions, which can limit the representation of higher-order dependencies that arise in biological sequences. We propose a higher-order attention mechanism that captures multi-token interactions beyond standard dot-product attention while remaining scalable through structured sparsity. Building on this mechanism, we introduce six Transformer architectures that span local-to-global dependency regimes and support both classification and regression settings. To enable controlled comparisons, we implement a unified, modular Transformer pipeline where attention is a plug-and-play component, allowing existing and custom attention modules to be evaluated under an identical training and evaluation protocol. Across TAPE protein benchmarks, our models achieve consistent improvements over strong Transformer baselines, with the largest gains on tasks requiring richer contextual interactions, while maintaining competitive efficiency.

1 Introduction

Biological sequences encode functional behaviour through dependencies that extend beyond simple pairwise relations. Across proteins, RNAs, and genomic regulatory regions, coordinated patterns of variation and structure frequently arise from the joint contribution of three or more positions. These higher order interactions influence stability, catalytic function, evolutionary adaptation, and regulatory control, and have been documented in a wide range of molecular systems.

In proteins, three way epistasis plays a central role in shaping structural and functional outcomes. Classical catalytic triads demonstrate that enzymatic activity depends on a coordinated triplet rather than independent residue pairs (**?**). Deep mutational scanning has shown that many fitness effects arise from combinations of three interacting residues, indicating that higher order couplings are pervasive in protein evolution (**??**). Coevolutionary analyses further reveal that residue triplets can mediate long range allosteric communication and cooperative stabilisation of tertiary structures (**??**). These findings cannot be explained by pairwise statistics alone.

Triplet interactions are also common in viral and microbial genomes. In rapidly evolving pathogens such as influenza and SARS–CoV–2, the functional effect of a mutation often depends on the presence of two additional mutations, forming synergistic sets that modulate receptor binding, antigenicity, or immune escape (**??**). Coordinated mutation triplets in HIV shape tropism and drug susceptibility (**?**).

Foundation Models for Science Workshop,38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Similar patterns appear in bacteria, where three way epistasis guides antibiotic resistance trajectories and compensatory adaptation (**?**).

Regulatory DNA and RNA sequences also rely on cooperative multi positional logic. Enhancer activity frequently depends on specific combinations of transcription factor binding sites whose effect emerges only when all required motifs appear in the correct arrangement (**??**). RNA structures contain coordinated nucleotide triplets that stabilise local folds, pseudoknots, or catalytic centres, and the phenotypic effect of a substitution often depends on the joint state of several additional nucleotides (**?**). These examples reinforce that combinatorial dependencies are a defining feature of non coding regions.

The prevalence of higher order dependencies has direct statistical implications. From a functional ANOVA perspective, any sequence–to–function mapping $f(x_1, \ldots, x_N)$ admits a decomposition

$$f = \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \sum_{i<j<k} f_{ijk}(x_i, x_j, x_k) + \cdots,$$

where f_{ijk} represents the pure three way interaction that remains after all main and pairwise contributions are removed. Conventional attention mechanisms approximate only the first two components through their bilinear structure, leaving f_{ijk} to be reconstructed indirectly through depth and nonlinearities, which is inefficient when higher order terms are substantial.

Information theory arrives at the same conclusion. True triplet dependence corresponds to nonzero interaction information,

$$I(X_i; X_j; X_k) = I(X_i; X_j) - I(X_i; X_j \mid X_k),$$

which quantifies whether the relationship between two positions changes once a third is taken into account. Biological sequences routinely display significant interaction information across triplets, meaning that models restricted to pairwise terms necessarily fail to extract all the information present in the data. Introducing an explicit third order interaction term therefore provides a principled way to capture structure that pairwise attention cannot represent.

{Maybe to be included: Finally, the third order component introduced in our attention mechanism can be understood as a structured analogue of a third order Volterra kernel or cubic polynomial transform in the latent representation. Classical systems theory shows that including explicit second and third order kernels increases the ability of a model to approximate nonlinear functionals with far fewer parameters than attempting to emulate the same interactions through compositions of linear operations. In this sense, the three way attention term provides a principled inductive bias that directs the model to search for cubic interaction patterns between embedding dimensions at three distinct sequence positions. This paragraph can be later linked to LoRA-like method of adding the 3D method on top of 2d model to improve performance instead of increasing the number of parameters in the dot-product attention. this can have the promise of instead of increasing the number of parameters, we are changing the way the model is extracting information. }

Taken together, these biological and statistical considerations highlight a recurring principle: many meaningful sequence features arise from higher order dependencies that cannot be represented adequately by pairwise models. Standard self attention mechanisms rely on a bilinear comparison between two positions and therefore lack an explicit mechanism for capturing coordinated triplets. This limits their ability to detect multi residue stabilisation motifs, epistatic mutation combinations, and regulatory triplets that jointly determine functional behaviour.

To address this gap, we introduce an attention mechanism that incorporates both a conventional pairwise term and an explicit third order interaction term capable of assigning importance to triplets of positions. This formulation directly targets the higher order components of the functional decomposition and provides the capacity to represent nonzero interaction information. Applied to the TAPE secondary structure prediction benchmark, the model consistently outperforms the standard two dimensional attention across multiple test conditions, demonstrating that explicit modelling of three way interactions yields measurable benefits in tasks where biological function is shaped by coordinated dependencies among multiple positions.

1.1 Summary of findings

2 Related Work

3 Method

Confidential

4 Experiments

We evaluate HOMA against strong Transformer baselines on the Tasks Assessing Protein Embeddings (TAPE) benchmark (**?**). Our study spans both residue-level and sequence-level prediction: we train all models from scratch on Secondary Structure (SS3/Q3; per-residue classification) and on two global regression tasks, Fluorescence and Stability. Unless stated otherwise, all experiments share the same data preprocessing, tokenization, training loop, optimizer and schedule, batch size, and training budget; When applicable, we use early stopping based on validation performance, using fixed random-seed controls for reproducibility.

4.1 Controlled Comparison and Implementation

To isolate the effect of attention design, we implement a unified Transformer pipeline in which the attention module is the only interchangeable component. All other architectural and training choices are held constant across models, and we match overall model scale as closely as possible (Table **??**). HOMA introduces additional parameters through the utility matrix U used in the higher-order interaction pathway; to prevent parameter count from becoming a confounder, we apply a low-rank factorization of U and analyze the effect of rank in Section 6. GPU A100 80 GB.

4.2 Datasets, Tasks, and Metrics

We use the official TAPE releases and evaluation protocols (**?**). For Secondary Structure, the model predicts one of three labels (helix/strand/coil) for each residue given the primary sequence; we train on the provided splits and evaluate on the standard test sets CB513, CASP12, and TS115, reporting Q3 accuracy and F1. Fluorescence and Stability are sequence-level regression tasks defined over mutated protein variants; for both tasks we report Spearman rank correlation (ρ) between predictions and measured values, consistent with TAPE. All tasks use residue-level tokenization with the standard IUPAC amino-acid alphabet (single-letter codes mapped to integer IDs), and we adopt a single sequence-length handling policy (padding/truncation with masking) across all models. Flurence leng sequences is small 240 max 237 padded to 240.

4.3 Baselines and Model Configurations

All baselines share the same Transformer backbone (embedding size, depth, head count, MLP width, normalization, and dropout), differing only in the attention mechanism. We include: **(i) global pairwise attention** (vanilla multi-head self-attention, MHSA) as the standard Transformer baseline; **(ii) block-wise/local attention** implemented with overlapping blocks to enforce locality and reduce quadratic cost, inspired by prior block-sparse and sliding-window attention designs; and **(iii) linear attention** based on low-rank / projection-style approximations (e.g., Linformer-type mechanisms) to improve efficiency at long sequence lengths. HOMA replaces the pairwise attention core with a higher-order mixed attention module while keeping the surrounding pipeline unchanged. Shared hyperparameters, parameter counts, and any task-specific settings are summarized in Table 1.

5 Results

We compare HOMA against the baseline attention mechanisms on Secondary Structure (SS3/Q3), Fluorescence, and Stability, training all models from scratch under the controlled pipeline described above. Table **??** reports the full results and parameter counts. Overall, HOMA improves consistently

over global pairwise attention and efficient-attention baselines, indicating that explicitly modeling higher-order interactions yields stronger protein representations under matched training conditions.

Across tasks, we observe complementary behavior depending on the local window size $w \in \{3, 5, 7\}$ used by HOMA. On **Secondary Structure** (CASP12), HOMA achieves the best performance among compared models; in particular, $w = 5$ yields **0.6588** Q3 accuracy and **0.6338** F1, outperforming the strongest baseline (Blockwise-2D) in both metrics. For **Fluorescence**, the best configuration shifts toward a larger context: $w = 7$ achieves the highest correlation, $\rho = 0.7388$, improving over Blockwise-2D ($\rho = 0.6998$) and suggesting that broader interaction coverage benefits sequence-level functional prediction. On **Stability**, HOMA again provides substantial gains over baselines, with $w = 5$ achieving the strongest score in the main comparison ($\rho = 0.7152$), improving notably over Blockwise-2D ($\rho = 0.6509$). Taken together, these results show that HOMA consistently improves both residue-level structural modeling and global property prediction, while window size trades off computational cost and effective receptive field: smaller windows are more efficient and remain competitive, whereas larger windows can further benefit sequence-level regression, especially on fluorescence. add the outperformance on the tape dataset while comparing it with the leader board and model specifics of the TAPE transformer.

Model	**CASP12**		**Fluorescence**	**Stability**
	F1	Acc	ρ	ρ
(A) Baseline Models				
Pairwise-2D	0.5257	0.5582	0.6647	0.6221
Blockwise-2D	0.6228	0.6368	0.6998	0.6509
Linear-2D	0.5466	0.5458	0.6821	0.5439
(B) Augmented Models				
HOMA ($w = 3$)	0.6226	0.6422	0.7016	0.6646
HOMA ($w = 5$)	**0.6338**	**0.6588**	0.7116	**0.7152**
HOMA ($w = 7$)	0.6289	0.6514	**0.7388**	0.7106

Table 1: T

able 2.Performance Report of Transformers with different attention modules.

6 Ablations and Analysis

We perform targeted ablation studies to isolate the contribution of three design and training choices in HOMA: (i) the low-rank factorization used to parameterize the U-matrix, (ii) initialization and optimization strategies for the pairwise (2D) attention backbone via pretraining and freezing, and (iii) the maximum sequence length used during training. All ablations are conducted on the TAPE secondary structure task, and we report accuracy@3 on the CASP12, TS115, and CB513 test sets. Figure 1 summarizes the resulting trends across datasets.

Rank ablation (Fig. 3A). We vary the rank used to factorize the U-matrix and observe a consistent advantage for the full-rank parameterization. Training from scratch with full U attains 66.6% on CASP12, 68.2% on TS115, and 65.5% on CB513, while the best low-rank setting in each case reaches 65.1% (rank-8) on CASP12, 67.9% (rank-8) on TS115, and 65.0% (rank-8) on CB513. On CASP12, accuracy decreases monotonically as rank is reduced from 128 to 8 (65.6% $\rightarrow$ 65.4% $\rightarrow$ 65.1%), yielding a 1.5-point drop relative to full-rank. In contrast, TS115 and CB513 are less sensitive to aggressive factorization: rank-8 remains within 0.3 and 0.5 points of full-rank, respectively, and slightly exceeds rank-32 and rank-128 (TS115: 67.9% vs. 67.7%/67.7%; CB513: 65.0% vs. 64.8%/64.4%). This pattern suggests that while CASP12 benefits from higher-rank capacity, a compact rank-8 approximation can capture most of the useful structure on TS115 and CB513.

Pretraining and freezing of pairwise attention (Fig. 3B). We next ablate the training procedure by varying how attention parameters are initialized and optimized. In the first setting, both pairwise and triadic attention parameters are randomly initialized and trained end-to-end within HOMA (2D from scratch). In the second setting, we first train a model with pairwise (2D) attention, then transfer the learned pairwise attention parameters into HOMA as initialization; we either continue updating these transferred parameters during HOMA training (pretrained, no freeze) or keep them fixed to reduce compute and memory overhead (pretrained, frozen). Figure 3B shows that freezing the transferred pairwise weights consistently underperforms the other strategies, yielding 64.5% on CASP12, 65.5% on TS115, and 62.6% on CB513. Allowing the transferred weights to continue training largely closes the gap to training from scratch: on CASP12, HOMA with 2D from scratch achieves 65.9% compared to 65.1% for pretrained, no freeze (a 0.8–0.9 point advantage for scratch), while on TS115 the two are nearly identical (68.0% vs. 67.9%). On CB513, however, pretrained initialization with continued training for pairwise attention weights provides a clear benefit, reaching 68.0% versus 65.0% for training from scratch (a 3.0-point improvement). Overall, these results indicate that transferred pairwise attention can serve as an effective initialization for HOMA when it is allowed to adapt during training, whereas freezing the transferred parameters limits adaptation to the triadic pathway and leads to systematic accuracy drops despite potential efficiency gains.

Effect of maximum sequence length (Fig. 3C). Our third ablation study examines how the maximum sequence length used during training affects secondary-structure accuracy. We compare training with sequences truncated/padded to lengths 256, 512, and 1024. Across all three test sets, using length 256 yields the lowest performance, consistent with information loss from aggressive truncation (63.5% on CASP12, 66.3% on TS115, and 63.6% on CB513). Increasing the limit to length 512 provides the best overall results, reaching 65.1% on CASP12, 67.9% on TS115, and 68.0% on CB513. Extending to length 1024 improves over length 256 but does not match length 512, with accuracies of 62.9% on CASP12, 67.2% on TS115, and 64.7% on CB513. These trends suggest a trade-off: allowing moderately longer contexts reduces truncation-induced information loss, but very long sequences can introduce additional irrelevant context or noise during optimization, which may partially offset the benefits of retaining more residues.

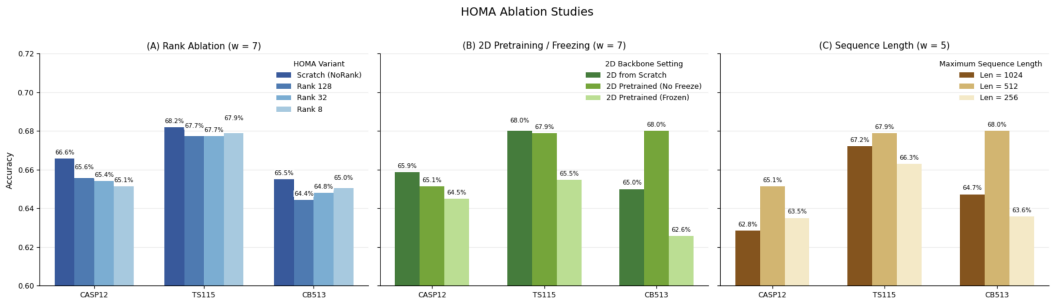

Figure 1: **HOMA ablations across datasets.** (A) Effect of low-rank setting (NoRank vs ranks 128/32/8) under $w = 7$. (B) Effect of 2D backbone initialization and freezing under $w = 7$. (C) Effect of maximum sequence length under $w = 5$. Bars report accuracy on CASP12, TS115, and CB513; y-axis is shared across panels for comparability.

7 Results

homa was tested with three different sliding window sizes 3,5 ,7. the datasets we tested against the TAPE () benchmarks. The baseline models the pairwise 2D attention and the multilped attention are recunstructed and trained from scratch on the three datasets Secondary structure, Flurescence and stability datasets more information on the model's congigurations can be found at table 1. HOMA seems to outperform the baseline models by more than 12% achieving an accuracy @ 3 of 65.88% on the secondary structure dataset. on the flurencne dataset HOMA outperforms the blockwise attention by 4% correlation score rho and on the Stability dataset our model outperforms the baselines by more than 6%. Note that our HOMA architecture with25.9M parameters outperformed the learning transformer model on the TAPE leaderboard. stability too? should we include the flurecence leader

Table 2: Model variants and configurations. SS denotes Secondary Structure; F&S denotes Fluorescence and Stability.

Model	Attention	Key hyperparameters	Params (SS)	Params (F&S)
(A) Baselines				
Pairwise-2D	Global MHSA	–	≈25.5M	≈20.9M
Blockwise-2D	Overlapping block MHSA	ℓ=b, s=15	≈25.5M	≈21.3M
Linear-2D	Linformer-style attention	k=50	≈26.1M	≈22.1M
(B) Proposed				
HOMA	Pairwise backbone + windowed Triple	ℓ=b, s=15, w=·, r=8	≈25.9M	≈21.5M
(C) Shared settings				
d_{model}	SS / F&S		512	256
Layers	SS / F&S		12	12
Heads	SS / F&S		8	8
FFN dim	SS / F&S		1024	128
Dropout	SS / F&S		0.4	0.4
LR	SS / F&S		10^{-4}	5×10^{-5}

board?
where to mention the table with the fairness of its study? then give the table. ...

8 Limitations and Broader Considerations

although performance is imporoved, there are considerations around compute and memory efficiency that can be of concern and can be further improved. optimization methods like FASTAttention can be rewritten for HOMA attention to improve the efifciency. should be applied to other fields and domains see how applicable it is acrooss different fields. how could it be heplful for CV or NLP?

training with more gpu doing pretraining on larger corpus and then finetuning. Other domains seem to be promising including image analysis. ...

9 Conclusion

...

Please read the instructions below carefully and follow them faithfully.

10 Submission of papers to NeurIPS 2024 Foundation Models for Science Workshop

10.1 Style

Papers to be submitted to NeurIPS 2024 must be prepared according to the instructions presented here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only acknowledgments and references* are allowed. Papers that exceed the page limit will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2024 are the same as those in previous years.

Authors are required to use the NeurIPS LaTeX style files obtainable at the NeurIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

10.2 Retrieval of style files

The style files for NeurIPS and other conference information are available on the website at

`http://www.neurips.cc/`

The file `neurips_2024.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

The only supported style file for NeurIPS 2024 is `neurips_2024.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy, `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

Preprint option If you wish to post a preprint of your work online, e.g., on arXiv, using the NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your work with the text "Preprint. Work in progress." in the footer. This version may be distributed as you see fit, as long as you do not say which conference it was submitted to. Please **do not** use the `final` option, which should **only** be used for papers accepted to NeurIPS.

At submission time, please omit the `final` and `preprint` options. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `neurips_2024.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 11, 12, and 13 below.

11 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 13 regarding figures, tables, acknowledgments, and references.

12 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

12.1 Headings: second level

Second-level headings should be in 10-point type.

12.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

13 Citations, figures, tables, references

These instructions apply to everyone.

13.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2024` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2024}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous" and include a copy of the anonymized paper in the supplementary material.

13.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

13.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

[1] Sample of the first footnote.

[2] As in this example.

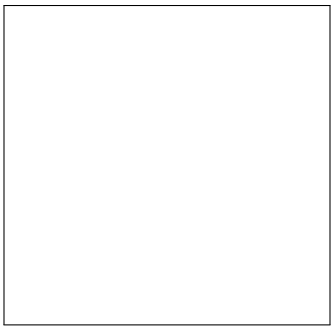

Figure 2: Sample figure caption.

Table 3: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

13.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 3.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 3.

13.5 Math

Note that display math in bare TeX commands will not create correct line numbers for submission. Please use LaTeX (or AMSTeX) commands for unnumbered display math. (You really shouldn't be using $$ anyway; see `https://tex.stackexchange.com/questions/503/why-is-preferable-to` and `https://tex.stackexchange.com/questions/40492/what-are-the-differences-between-align-equation-and-displaymath` for more information.)

13.6 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

14 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

14.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2024/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to automatically hide this section in the anonymized submission.

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

A Appendix / supplemental material

Optionally include supplemental material (complete proofs, additional experiments and plots) in appendix. All such materials **SHOULD be included in the main submission.**