Hybrid Attention State Space Models for Symbolic Calculation of Squared Amplitudes

Karaka Prasanth Naidu
Department of Electrical Engineering
Indian Institute of Technology Dhanbad
22je0456@iitism.ac.in

Eric A. F. Reinhardt
Department of Physics and Astronomy
University of Alabama
eareinhardt@crimson.ua.edu

Victor Baules
Department of Physics and Astronomy
University of Alabama
vabaules@crimson.ua.edu

Nobuchika Okada
Department of Physics and Astronomy
University of Alabama
okadan@ua.edu

Sergei Gleyzer
Department of Physics and Astronomy
University of Alabama
sgleyzer@ua.edu

Abstract

Calculating squared amplitudes is a key step in computing cross sections needed to compare experimental data with theoretical predictions. However, mapping amplitude expressions to their squared amplitude expressions is computationally expensive. Prior works have formulated this task as a sequence-to-sequence problem, demonstrating the effectiveness of Transformer-based encoder–decoder architectures. Despite these successes, such approaches have been limited to relatively short sequences and fail to scale effectively to longer inputs, primarily due to the limitations of attention mechanisms in handling extended context windows. State Space Models (SSMs), such as Mamba, offer a recurrent alternative that can achieve performance comparable to that of Transformers on some tasks. In this work, we investigate hybrid Attention and SSM architectures and show that they outperform Vanilla Transformers in a low-data, long-sequence task. The hybrid Attention-SSMs achieving up to a $\sim$4% improvement in token accuracy and $\sim$40% improvement in full sequence accuracy on the task of calculating squared amplitudes for electroweak physics processes.

1 Introduction

Calculating cross sections is a critical step in comparing experimental observations with physics theory. In the case of high-energy physics, one part of this calculation is to determine the probabilities of interactions occurring. To perform this step, the standard model of particle physics uses a wave function representation of particles to describe interaction processes with an expression known as an amplitude. Using an amplitude, the corresponding squared amplitude provides the probability density for a process.

Determining the amplitude of the interaction requires relatively little calculation. However, mapping amplitude expressions to squared amplitude expressions can be computationally expensive and time consuming [AGP23]. The work in [AGP23] explores the problem as a sequence-to-sequence

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

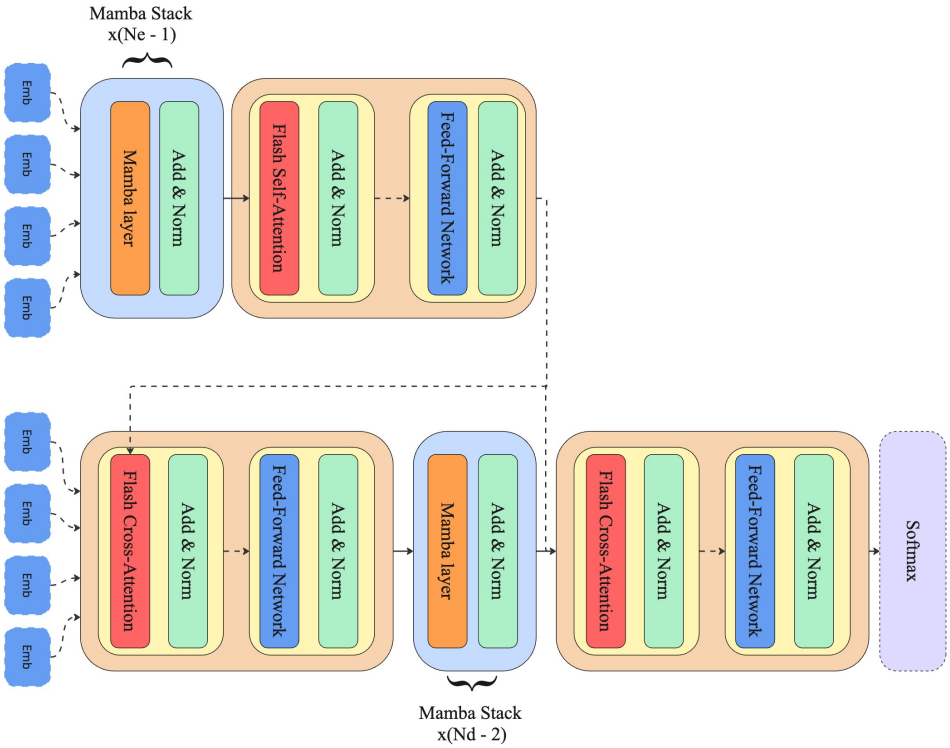

Figure 1: Overview of our model architecture.

mapping and showed the efficacy of Transformer-based models [Vas+23]. The work in [Bha+24] used transformer models with SineKAN [RRG25] layers at the end.

Works like [GD23] show that State Space Models (SSMs) perform comparably to transformers on various tasks with fewer parameters and have better long-context capability. Previous work on SSMs ([VAM25]) showed that SSMs are sample efficient, i.e. requiring less data to perform well that might be required by transformer models. These advantages motivate the exploration SSMs for squared amplitude calculation. In this work, we explore Hybrid SSMs, where we use both Mamba layers and transformer attention layers, more information about model architecture in section 2.2.

2 Model Description

2.1 MAMBA State Space Model

State Space Models introduced in [Gu+20] provide an alternative sequence mixing mechanism which processes sequences through a recurrence relation. Continous Latent State Space Models map 1-D function or sequence $x(t) \in \mathbb{R} \rightarrow y(t) \in \mathbb{R}$ through $h(t) \in \mathbb{R}^N$ via linear ordinary differential equations as follows:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}x(t) \tag{1}$$

$$y(t) = \mathbf{C}h(t) \tag{2}$$

A Discrete SSM is obtained after applying Zero-Order Hold(ZOH) and is defined as follows:

$$H_i = \mathbf{A}H_{i-1} + \mathbf{b}x_i^\top \tag{3}$$

$$y_i = H_i^\top \mathbf{c} \tag{4}$$

This model is input-independent, as the same parameters are applied for both relevant and irrelevant features. Consequently it lacks the ablity to reset or overwrite its hidden states. To introduce feature dependence, Mamba [GD23] employs a learnable linear projection layer over x prior to

discretization, serving as a selection mechanism. This ensures that the parameters become effectively feature-dependent. The resulting model is given by:

$$H_i = A_i \odot H_{i-1} + B_i \odot X_i \tag{5}$$

$$y_i = H_i^\top c_i \tag{6}$$

where $X_i = \mathbf{1}_r x_i^\top \in \mathbb{R}^{r \times d}$ is an r-sized stack of the input, $A_i \in \mathbb{R}^{r \times d}$ represents d diagonal matrices of size $r \times r$, $B_i \in \mathbb{R}^{r \times d}$, $c_i \in \mathbb{R}^r$, and $\odot$ is the Hadamard product. Mamba defines A_i and B_i with a shape of $(\ldots, d)$, allowing for unique parameters in each hidden dimension making Mamba more expressive than previous variant S4 [GGR22] and H3 [Fu+23].

2.2 Hybrid Variants

Previous works like [Var+23], [Fu+23] show that adding attention to SSMs lead to strong performance in language modeling. In this work we explore scaling Encoder-Decoder variant introduced in [Pit+24] and [Var+23] where a vanilla transformer encoder-decoder architecture is used but replacing all Self-Attention components with Mamba layers with cross attention kept intact. The specific architecture used, shown in fig. 1, was obtained after experimentation with different configurations of Attention, Multilayer Perceptron (MLP) and SSM components. The addition of Attention components leads to an additional $O(n^2)$ for training and $O(n)$ for autoregressive inference.

2.3 Dataset

Data are generated using the MARTY software package [UMA21]. Inputs to the model are the amplitudes of the Feynman diagrams, and labels are the squared amplitudes. All diagrams are restricted to include only on-shell final-state particles. They are further restricted to not include permutations of particle ordering, significantly limiting the total amount of available data. This differs from [Bha+24] and [AGP23] where permutations were included. Data was generated for electroweak, quantum electrodynamics, and quantum chromodynamics models. The data representation is in text form using standard notation from high-energy physics with scalar pre-factors, masses, and momenta, and with vectors, matrices, and tensors represented using index values. An example amplitude sequence for electron-electron scattering is given in fig. 2 and the associated squared amplitude is given in fig. 3.

```
-1/2*i*e^2*gamma_{+%\sigma_165,%gam_145,%gam_146}*gamma_{%
\sigma_165,%gam_147,%del_137}*e_{i_3,%gam_146}
(p_1)_u*e_{k_3,%del_137}(p_2)_u*e_{l_3,%gam_145}
(p_3)_u^(*)*e_{i_5,%gam_147}(p_4)_u^(*)/(m_e^2 + -s_13 + 1/2*reg_prop)
```

Figure 2: Amplitude of $e^-e^- \to e^-e^-$ scattering.

```
2*e^4*(m_e^4 + -1/2*m_e^2*s_13 + 1/2*s_14*s_23 + -1/2*m_e^2*s_24 +
1/2*s_12*s_34)*(m_e^2 + -s_13 + 1/2*reg_prop)^(-2)
```

Figure 3: Squared amplitude of $e^-e^- \to e^-e^-$ scattering.

The amplitude and squared amplitude expressions are then tokenized as the separable components mentioned as the standard notation. The indices generated by MARTY have no enforced order and accumulate during the data generation process. To preserve the meaning of these indices, they are replaced by

```
INDEX_0, INDEX_1,...
```

and are normalized to start from 0 going from left to right with tokens being recycled on repeated appearance of those indices. Electroweak 2-2 (EW 2-2) contains a total of 8512 rows, and Electroweak 2-3 (EW 2-3) consists of 110635 rows.

Type	Amplitude token limit	Squared Amplitude token limit	Number of rows
EW 2-2	512	512	6235
	512	1024	7890
	512	2048	7918
EW 2-3	512	512	47451
	512	1024	62730
	512	2048	74567

Table 1: Number of data rows in each configuration

3 Results and Discussion

For all experiments, a 3-layer encoder and a 4-layer decoder are used. We count a pair of Mamba kernels as a layer and a Self/Cross Attention layer with FFN as a layer with each feedforward dimensions set to 512. All models were trained close to convergence using a learning rate that decayed linearly across epochs. We used a warm-up phase for 10% of the training period and gradient norms were clipped to a unit value which helped stabilize training and improve convergence, particularly in the early stages. All Transformer components employ rotary positional embeddings (RoPE), which provided better generalization for longer contexts than absolute positional encodings in preliminary validation.

To understand the scaling nature of our model, we have performed training and evaluation on different data settings defined by a maximum token limit in source and target expressions. We first split the data and then filter the datasets by token length. Models are then evaluated based on sequence accuracy, where only exact matches between the generated and original sequences are considered. Testing is performed on an unseen test set comprising approximately 5–10% of the training data for each process.

Our current evaluation focuses on electroweak (EW) processes, which provide a controlled and well-understood benchmark for symbolic amplitude–squared mapping. Quantum chromodynamics (QCD) processes were excluded from this study due to the limited size of available datasets, which were insufficient for stable model training. Extending this approach to larger QCD datasets once available will be an important next step. Future work will also investigate whether the hybrid Attention SSM architecture can capture broader high-energy physics symbolic computations.

Note: We used embedding size of 512, 8 attention heads and our model has **14.6M** parameters wheras transformer model has 15.4M parameters.

Table 2: Token Accuracy across different configurations.

Process Type	amp_max_len	sqamp_max_len	Transformer	Ours
EW 2-2	512	512	95.78%	99.98%
	512	1024	88.90%	99.97%
	512	2048	88.65%	99.95%
EW 2-3	512	512	96.11%	99.98%
	512	1024	89.68%	99.97%
	512	2048	76.48%	99.96%

In a task where sequence accuracy is very sensitive to token accuracy, the hybrid Attention-SSM performs well. By scaling the max token threshold in dataset, we observed a drop in the performance helping us understand the limitations of these approaches for extremely long sequences. We show that even in longer sequences like 2048 tokens, the hybrid Attention-SSM retains performance. From the results presented in table 3 it is evident that adding attention components to State Space Models helps achieve better performance than transformer models with fewer learnable parameters.

Table 3: Sequence Accuracy across different configurations.

Process Type	amp_max_len	sqamp_max_len	Transformer	Ours	Data Retention
EW 2-2	512	512	43.25%	96.47%	73.25%
	512	1024	29.66%	95.36%	92.70%
	512	2048	30.03%	91.00%	93.03%
EW 2-3	512	512	55.56%	97.35%	42.89%
	512	1024	51%	95.39%	56.70%
	512	2048	28%	92.23%	67.40%

4 Conclusion and Future Work

In this work, we present a new solution to the task of solving squared amplitudes for amplitudes in particle physics. Previous works have explored Transformer models for this task, approaching it as a sequence-to-sequence translation. We show that a hybrid architecture combining Mamba SSMs with Attention mechanisms can improve performance on low-data, large sequence scenarios. The hybrid Attention-SSMs achieve $\sim$4% higher token accuracy and related $\sim$40% higher sequence compared to Vanilla Transformers. Future works could explore these hybrid Attention-SSMs for additional physics models and more general seq-to-seq tasks.

References

[Gu+20] Albert Gu et al. *HiPPO: Recurrent Memory with Optimal Polynomial Projections*. 2020. arXiv: 2008.07669 [cs.LG]. URL: https://arxiv.org/abs/2008.07669.

[UMA21] Grégoire Uhlrich, Farvah Mahmoudi, and Alexandre Arbey. "– Modern ARtificial Theoretical phYsicist A C++ framework automating theoretical calculations Beyond the Standard Model". In: *Computer Physics Communications* 264 (2021), p. 107928. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2021.107928. URL: https://www.sciencedirect.com/science/article/pii/S001046552100062X.

[GGR22] Albert Gu, Karan Goel, and Christopher Ré. "Efficiently Modeling Long Sequences with Structured State Spaces". In: *The International Conference on Learning Representations (ICLR)*. 2022.

[AGP23] Abdulhakim Alnuqaydan, Sergei Gleyzer, and Harrison Prosper. "SYMBA: symbolic computation of squared amplitudes in high energy physics with machine learning". In: *Machine Learning: Science and Technology* 4.1 (Jan. 2023), p. 015007. ISSN: 2632-2153. DOI: 10.1088/2632-2153/acb2b2. URL: http://dx.doi.org/10.1088/2632-2153/acb2b2.

[Fu+23] Daniel Y. Fu et al. *Hungry Hungry Hippos: Towards Language Modeling with State Space Models*. 2023. arXiv: 2212.14052 [cs.LG]. URL: https://arxiv.org/abs/2212.14052.

[GD23] Albert Gu and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *arXiv preprint arXiv:2312.00752* (2023).

[Var+23] Ali Vardasbi et al. "State Spaces Aren't Enough: Machine Translation Needs Attention". In: *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*. Ed. by Mary Nurminen et al. Tampere, Finland: European Association for Machine Translation, June 2023, pp. 205–216. URL: https://aclanthology.org/2023.eamt-1.20/.

[Vas+23] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[Bha+24] Ritesh Bhalerao et al. "S-KANformer: Enhancing Transformers for Symbolic Calculations in High Energy Physics". In: *Machine Learning and the Physical Sciences Workshop at NeurIPS 2024*. 2024. URL: https://ml4physicalsciences.github.io/2024/files/NeurIPS_ML4PS_2024_118.pdf.

[Pit+24] Hugo Pitorro et al. *How Effective are State Space Models for Machine Translation?* 2024. arXiv: 2407.05489 [cs.CL]. URL: https://arxiv.org/abs/2407.05489.

[RRG25] Eric Reinhardt, Dinesh Ramakrishnan, and Sergei Gleyzer. "SineKAN: Kolmogorov-Arnold Networks using sinusoidal activation functions". In: *Frontiers in Artificial Intelligence* 7 (Jan. 2025). ISSN: 2624-8212. DOI: 10.3389/frai.2024.1462952. URL: http://dx.doi.org/10.3389/frai.2024.1462952.

[VAM25] Shweta Verma, Abhinav Anand, and Mira Mezini. *CodeSSM: Towards State Space Models for Code Understanding*. 2025. arXiv: 2505.01475 [cs.SE]. URL: https://arxiv.org/abs/2505.01475.

A Technical Appendices and Supplementary Material

A.1 Repository

The code for our work with additional details regarding the experiments is available at the following Github repository: https://github.com/prasanth30/SYMBA_SSM/tree/endterm