Quantum-aware Transformer model for state classification

Przemysław Sekuła[1,2][0000−0002−4599−1077], Michał Romaszewski[1][0000−0002−8227−929X], Przemysław Głomb[1][0000−0002−0215−4674], Michał Cholewa[1][0000−0001−6549−1590], and Łukasz Pawela[1][0000−0002−0476−7132]

[1] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka 5, 44-100 Gliwice, Poland
[2] University of Maryland, College Park, Department of Civil and Environmental Engineering, MD, USA
{psekula,mromaszewski,przemg,mcholewa,lpawela}@iitis.pl

Abstract. Entanglement is a fundamental feature of quantum mechanics, playing a crucial role in quantum information processing. However, classifying entangled states, particularly in the mixed-state regime, remains a challenging problem, especially as system dimensions increase. In this work, we focus on bipartite quantum states and present a data-driven approach to entanglement classification using transformer-based neural networks. Our dataset consists of a diverse set of bipartite states, including pure separable states, Werner entangled states, general entangled states, and maximally entangled states. We pretrain the transformer in an unsupervised fashion by masking elements of vectorized Hermitian matrix representations of quantum states, allowing the model to learn structural properties of quantum density matrices. This approach enables the model to generalize entanglement characteristics across different classes of states. Once trained, our method achieves near-perfect classification accuracy, effectively distinguishing between separable and entangled states. Compared to previous Machine Learning, our method successfully adapts transformers for quantum state analysis, demonstrating their ability to systematically identify entanglement in bipartite systems. These results highlight the potential of modern machine learning techniques in automating entanglement detection and classification, bridging the gap between quantum information theory and artificial intelligence.

Keywords: Quantum entanglement · State classification · Transformers · Large language models

1 Introduction

The entanglement phenomenon is at the heart of quantum information, enabling its key applications such as quantum computing, secure communication, and enhanced metrology. Unlike classical correlations, entanglement represents a uniquely quantum feature where the state of a system cannot be described

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

independently of its subsystems. This fundamental property underlies protocols like quantum teleportation, superdense coding, and quantum key distribution, as well as quantum computational speedups [25,17]. However, not all quantum states exhibit entanglement, and distinguishing entangled states from separable ones is a crucial yet challenging problem in quantum information science. The ability to efficiently classify quantum states has direct implications for the practical implementation of quantum technologies, motivating the development of reliable entanglement detection and classification methods.

Bipartite quantum states, which describe systems naturally partitioned into two subsystems, are fundamental in quantum information science and serve as a basis for studying quantum correlations and computational advantages. These states reside in a tensor product space $\mathrm{L}(\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2})$, where entanglement emerges as a crucial resource for various quantum information applications [14,6], such as secure communication, quantum-enhanced computation, and efficient information transfer. For pure bipartite states, i.e., vectors in $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$, entanglement can be clearly identified: a state is separable if and only if it can be expressed as a tensor product of subsystem states. Any departure from this structure signifies entanglement, which directly influences quantum nonlocality, measurement-based quantum computing, and the efficiency of entanglement-assisted protocols.

The identification of entanglement for bipartite states is straightforward in the case of pure states but becomes significantly more challenging when considering mixed bipartite states. A variety of analytical and numerical methods have been developed for the characterization and detection of mixed-state entanglement. One of the most celebrated approaches is based on the *positivity of the partial transpose (PPT)*, introduced by Peres [19] and Horodecki [11]. While the PPT criterion is both necessary and sufficient for separability in low-dimensional cases ($\mathbb{C}^2 \otimes \mathbb{C}^2$ and $\mathbb{C}^2 \otimes \mathbb{C}^3$), in higher dimensions, a state can be *PPT and still entangled*. Such states are known as *bound entangled* states [12] (see Fig. 1 for sketch). They cannot be distilled into pure entangled states using local operations and classical communication (LOCC), rendering them entangled yet practically "inaccessible" for certain quantum information protocols [1]. This discovery highlighted the limitations of the PPT criterion and the intricate nature of entanglement in higher-dimensional systems. To emphasize this difference, entangled states which are not bound entangled are sometimes called free entangled states.

To address these limitations, other techniques—particularly *entanglement witnesses*—have been introduced [14,10]. An entanglement witness is a Hermitian operator W with the property that $\mathrm{Tr}(W\,\rho_{\mathrm{sep}}) \geq 0$ for all separable states ρ_{sep}, but $\mathrm{Tr}(W\,\rho_{\mathrm{ent}}) < 0$ for at least one entangled state ρ_{ent}. Finding and optimizing entanglement witnesses can often be formulated via semidefinite programming techniques, and in many cases, witnesses can be tailored to detect specific classes of entangled states, including those exhibiting bound entanglement. Additional approaches to mixed-state entanglement include various *entanglement measures* (e.g., negativity, entanglement of formation), which attempts to quantify the degree of entanglement in a given density operator [24,20].

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

Another class of approaches relies on machine learning (ML) to identify entanglement directly from the data. In [8], authors use automated ML for state classification. Instead of directly measuring entanglement properties, the state is reconstructed, and entanglement is inferred from the data itself. Recently, Transformers have also been applied to quantum random number validation [7], showcasing their ability to handle large input sequences efficiently and perform multiple statistical tests in parallel. The same self-attention mechanism that captures subtle global dependencies in random bit streams can likewise model the intricate correlations of bipartite quantum states, suggesting that Transformers are a promising architecture for entanglement classification under partial or noisy data.

Despite significant progress, a complete classification of bipartite entangled mixed states remains an open challenge, particularly as system dimensions grow. In this work, we take a data-driven approach to this problem by generating a diverse dataset of pure and mixed states through multiple methods and employing transformer-based neural networks to analyze their properties and classify them. We demonstrate that entanglement identification can be performed effectively from the data itself, extending the range of successful classification beyond previous studies. Furthermore, we validate the application of transformer architectures in this domain, achieving a breakthrough where prior deep learning approaches [8] have struggled. By integrating machine learning with established theoretical criteria, we provide a scalable framework for systematically detecting and categorizing entanglement, bridging the gap between quantum information theory and modern Artificial Intelligence techniques.

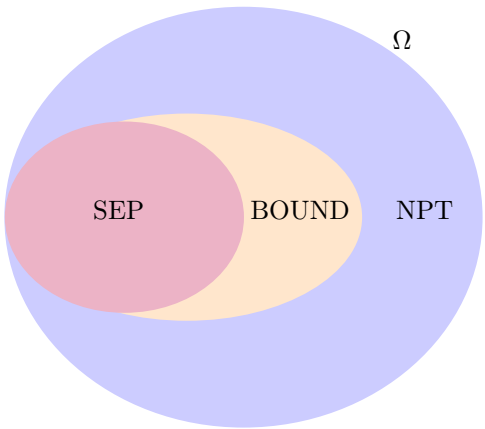

Fig. 1. Schematic representation of the set of mixed quantum states Ω depicting separable states (SEP) ρ_{sep}, bound entangled states (BOUND) σ_{bound} and negative partial transpose states (NPT) ξ_{npt}.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

This paper is organized as follows. In Section 2, we explain the methodology for constructing our dataset. In Section 3, we detail our proposition of the Quantum-aware Transformer model and its training scheme. In Section 4, we present the results of validation experiments and discussion.

The code for this paper is publicly available on GitHub[3] under an open license to facilitate result reproducibility and transparency and the data can be shared upon a reasonable request.

2 Dataset generation

Quantum states, the basic objects of quantum mechanics, can be broadly classified as *pure* or *mixed*. A *pure state* is described by a single d-dimensional vector $|\psi\rangle$ in a complex Euclidean space $\mathbb{C}^d$, and a corresponding density operator $\rho = |\psi\rangle\langle\psi|$. This operator satisfies $\rho^2 = \rho$ and $\mathrm{Tr}(\rho) = 1$. When $d = 2$, the corresponding system is called a qubit; for $d = 3$, it is called a qutrit.

In contrast, a *mixed state* is represented by a statistical ensemble of pure states, described by a density operator $L(\mathbb{C}^d) \ni \rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$, where $p_i \in [0, 1]$ and $\sum_i p_i = 1$ [16]. Such states often emerge from partial traces of larger systems or incomplete information about the quantum system under study. Note that ρ is a positive semidefinite matrix.

In many quantum information scenarios, one focuses on *bipartite* states. These describe a physical system that can be naturally partitioned into two subsystems, $|\psi\rangle \otimes |\phi\rangle$, associated with complex Euclidean spaces $\mathbb{C}^{d_1}$ and $\mathbb{C}^{d_2}$. The total state then lives in the tensor product space $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$. Bipartite quantum systems are not only a cornerstone for foundational studies of quantum correlations but also play a central role in key quantum information tasks such as quantum teleportation, quantum key distribution, and superdense coding.

For *pure bipartite states*, there exists a straightforward way to distinguish entangled from separable states: a pure bipartite state $|\psi\rangle \in \mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ is separable if and only if it can be written as $|\xi\rangle = |\psi\rangle \otimes |\phi\rangle$. Any deviation from this product structure indicates entanglement.

In this work, we study the following cases of discrimination between separable and entangled states:

1. two-qubit states, $\mathbb{C}^2 \otimes \mathbb{C}^2$,
2. qubit-qutrit systems, $\mathbb{C}^2 \otimes \mathbb{C}^3$,
3. qutrit-qutrit systems, $\mathbb{C}^3 \otimes \mathbb{C}^3$,
4. ququart-ququart systems, $\mathbb{C}^4 \otimes \mathbb{C}^4$,

For each of these, we generate a dataset consisting of:

1. pure separable states,
2. general entangled states,
3. Werner entangled states (for all but qubit-qutrit systems),
4. maximally entangled states (for all but qubit-qutrit systems),

[3] https://github.com/iitis/LQM

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

5. bound entangled states from the family by Horodecki [13] (for qutrit-qutrit systems).

For this work we assume we have access to the full tomography [3] of each ρ, hence we encode each state as a vector of $2d_1d_2$ real variables.

The details of the sampling are described in the following subsections. Uniform sampling of quantum states (either pure or mixed) was conducted utilizing the QuantumInformation.jl package [5].

2.1 Sampling pure separable states

In this case, we need to sample uniformly normalized vectors of the form $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2} \ni |\psi\rangle = |\phi_1\rangle \otimes |\phi_2\rangle$. This is done in the following steps:

1. Sample a non-normalized $|x\rangle \in \mathbb{C}^{d_1}$ with each element x_i such that $\mathrm{Re}(x_i) \sim N(0,1)$ and $\mathrm{Im}(x_i) \sim N(0,1)$.
2. Normalize $|x\rangle$: $|\phi_1\rangle = \frac{|x\rangle}{\||x\rangle\|}$.
3. Repeat steps 1 and 2 to obtain $|\phi_2\rangle$.
4. Put $|\psi\rangle = |\phi_1\rangle \otimes |\phi_2\rangle$.

This procedure ensures that each $|\phi_i\rangle$ is sampled from the Haar measure.

2.2 Sampling Werner entangled states

A Werner state is a mixed state having the form

$$\rho_{\mathrm{wer}} = (1-p)\,|\psi\rangle\langle\psi| + p\rho^*, \tag{1}$$

where ρ^* is the maximally mixed state, $\rho^* = \frac{\mathbb{1}}{d^2}$, and

$$|\psi\rangle = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |ii\rangle\,. \tag{2}$$

It can be shown that the state remain entangled for

$$p < \frac{d}{d+1}. \tag{3}$$

We sample p uniformly in this interval and construct ρ_{wer}.

2.3 Sampling general entangled states

We sample general entangled states by uniformly sampling the state of all mixed quantum states and only accepting the state as entangled when it is NPT. The steps are:

1. Sample a square Ginibre matrix, G, of dimension d_1d_2. Elements G_{ij} pertain to the complex normal distribution $\mathrm{Re}(G_{ij}) \sim N(0,1)$ and $\mathrm{Im}(G_{ij}) \sim N(0,1)$.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

2. Calculate $W = GG^\dagger$.
3. Normalize the trace, $\rho = \frac{W}{\mathrm{Tr}\, W}$.
4. Check the Peres-Horodecki criterion. If ρ is NPT, accept it into the set; otherwise, repeat the procedure.

Note that this procedure is quite efficient, especially as the dimension increases, as the relative volume of the separable states diminishes [26,27,28].

2.4 Sampling maximally entangled states

We sample maximally entangled states by sampling unitary matrices, vectorizing them, and renormalizing them [21]. The procedure is:

1. Sample a square Ginibre matrix G of dimension d as described in Section 2.3.
2. Calculate its QR decomposition $G = QR$, where Q is a unitary matrix and R is upper triangular.
3. Multiply i-th column of Q, Q_i, by the phase of the corresponding diagonal element of R, R_{ii}, thus obtaining the i-th column of a unitary matrix U. This step is necessary to ensure the proper distribution of eigenvalues of U [15,18].
4. Vectorize the matrix U and normalize by $\frac{1}{\sqrt{d}}$.

2.5 Sampling bound entangled states

This procedure is based on a family of bound entangled states described in [13]. First, let us introduce

$$\begin{aligned} \sigma_+ &= \frac{1}{3}\left(|01\rangle\langle 01| + |12\rangle\langle 12| + |20\rangle\langle 20|\right) \\ \sigma_- &= \frac{1}{3}\left(|10\rangle\langle 10| + |21\rangle\langle 21| + |02\rangle\langle 02|\right) \\ |\psi\rangle &= \frac{1}{\sqrt{3}}\left(|00\rangle + |11\rangle + |22\rangle\right). \end{aligned} \tag{4}$$

We construct a state ρ as follows

$$\rho_\alpha = \frac{2}{7}|\psi\rangle\langle\psi| + \frac{\alpha}{7}\sigma_+ + \frac{5-\alpha}{7}\sigma_-. \tag{5}$$

Depending on parameter α this state can be separable ($2 \leq \alpha \leq 3$), bound entangled ($3 < \alpha \leq 4$) or free entangled ($4 < \alpha \leq 5$). Hence, we sample α uniformly in the range $(3, 4]$.

The dataset construction procedure is explicitly designed to avoid introducing spurious correlations between the structural features of the density matrices and the generative mechanisms from which they are sampled. Each class of quantum states is synthesized using statistically independent and physically grounded protocols that are aligned with the operational definitions of separability and entanglement. Pure separable states are sampled from the product of independent

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

Haar measures on the unit spheres in $\mathbb{C}^{d_1}$ and $\mathbb{C}^{d_2}$, ensuring uniform coverage of the separable manifold. General entangled states are drawn from the space of density operators via normalized Wishart ensembles (Ginibre matrices) and are subsequently filtered through the Peres-Horodecki criterion to enforce negative partial transposition (NPT), thereby guaranteeing entanglement. Additionally, the dataset incorporates structured families of entangled states with analytically controlled properties: Werner states, defined by convex combinations of maximally entangled states and the maximally mixed state, and known to be entangled for $p < \frac{d}{d+1}$; and bound entangled states from the Horodecki construction, parameterized by $\alpha \in (3, 4]$, which guarantees PPT entanglement in the two-qutrit case. The inclusion of these analytically tractable families serves to increase the geometric and algebraic diversity of the entangled class, thereby mitigating the risk that the neural network overfits to artifacts of a single synthesis algorithm rather than learning entanglement-relevant features intrinsic to the density matrices themselves.

3 Quantum-aware Transformer model

We employ a Transformer-based architecture [23] to process quantum state matrices, leveraging its powerful self-attention mechanism to model complex relational structures within the data. While Transformers were originally developed for natural language processing, their ability to capture long-range dependencies and contextual interactions makes them well-suited for structured matrix data as well.

In our approach, we represent quantum state matrices as sequences of tokens, where each token encodes the real and imaginary components of a matrix element. This tokenized representation allows the Transformer to globally attend to all parts of the matrix, modeling both local and non-local correlations that are critical in quantum systems.

We further adopt a masked autoencoding training strategy, enabling the model to learn robust representations by reconstructing missing matrix elements. This encourages the network to internalize key quantum properties such as coherence and entanglement. As a result, the Transformer not only reconstructs structured quantum data effectively but also enhances performance in downstream tasks, such as entanglement classification.

The Transformer model was chosen for its ability to flexibly capture both pairwise and more complex interactions within data, without imposing assumptions about spatial locality or sequential order. Such assumptions, common in other architectures, could constrain the modeling of quantum data, where relevant relationships may span arbitrary positions in the data structure.

3.1 Model Description

We employ a masked Transformer architecture, referred to as the *MaskedTransformer*, to reconstruct partially masked 2D quantum data. The input is first

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

reshaped from $\left[B, 2N^2\right]$ into $\left[B, N^2, 2\right]$, where B stands for the size of data batch, N is the size of the original square matrix (each entry contains real and imaginary parts). Tokens in this representation are randomly masked and replaced by a *mask token*, and the partially masked sequence is passed through a Transformer encoder composed of multi-head self-attention layers and feed-forward networks.

Each token embedding is further augmented with a trainable positional vector $\mathbf{p}_i$, ensuring that the Transformer retains the relative location of each matrix element in the spatial grid. Concretely, we added $\mathbf{p}_i \in \mathbb{R}^d$ to the embedded token $\mathbf{x}_i$, yielding

$$\tilde{\mathbf{x}}_i \;=\; \mathrm{Embed}\big(\mathbf{x}_i\big) \;+\; \mathbf{p}_i, \tag{6}$$

where Embed is the token embedding function. This positional encoding is crucial to provide the necessary positional information for the self-attention mechanism.

A simple linear decoder projects the final encoded representations back to real and imaginary components, reconstructing both the originally masked and unmasked portions. This masked-reconstruction approach is analogous to masked autoencoders or BERT [4] like masking methods, thereby encouraging the model to infer missing entries from the surrounding context.

In our experiments, we tested three configurations corresponding to $n = 4$ (for $\mathbb{C}^2 \otimes \mathbb{C}^2$), $n = 6$ (for $\mathbb{C}^2 \otimes \mathbb{C}^3$), $n = 9$ (for $\mathbb{C}^3 \otimes \mathbb{C}^3$), and $n = 16$ (for $\mathbb{C}^4 \otimes \mathbb{C}^4$). All other Transformer hyperparameters (embedding dimension, number of attention heads, number of layers, and dropout) were selected during the initial research phase by trail-and-error method, and remained the same across these experiments.

We used separate datasets for each experiment for pretraining and classification, as presented in Table 1. In every experiment, the dataset was divided into training, validation, and test subsets. The validation evaluation was performed every epoch, and the test evaluation was performed after the training was completed.

3.2 Pretraining

In this stage, we train the *MaskedTransformer* from scratch as an autoencoder. Specifically, we:

- Use the datasets of 10 mln (for $\mathbb{C}^2 \otimes \mathbb{C}^2$), 16 mln ($\mathbb{C}^2 \otimes \mathbb{C}^3$ and $\mathbb{C}^3 \otimes \mathbb{C}^3$), and 18 mln (for $\mathbb{C}^4 \otimes \mathbb{C}^4$), described in detail in Table 1 in the *Pretraining* columns.
- Split the dataset it into training (90%), validation (5%), and test (5%) partitions.
- Randomly mask a given fraction (15%) of tokens, replacing them with a learned mask token.
- Pass the masked tokens through the Transformer encoder and reconstruct them via a linear decoder head.
- Optimize the full reconstruction loss using mean squared error (MSE).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

Group name	$\mathbb{C}^2 \otimes \mathbb{C}^2$	$\mathbb{C}^2 \otimes \mathbb{C}^3$	$\mathbb{C}^3 \otimes \mathbb{C}^3$	$\mathbb{C}^4 \otimes \mathbb{C}^4$
Pretraining				
sep	4,000,000	8,000,000	6,000,000	9,000,000
general-ent	2,000,000	8,000,000	2,000,000	3,000,000
werner-ent	2,000,000		2,000,000	3,000,000
max-ent	2,000,000		2,000,000	3,000,000
horodecki-bound			2,000,000	
horodecki-ent			2,000,000	
Classification				
sep	1,000,000	1,000,000	1,000,000	1,000,000
general-ent	300,000	1,000,000	500,000	300,000
werner-ent	300,000		500,000	300,000
max-ent	300,000		500,000	300,000
horodecki-bound			500,000	
horodecki-ent			500,000	

Table 1. Training data sizes for Pretraining and Classification tasks. Numbers indicate samples used for each model configuration and data type. Group names correspond to the type of data (pure separable, general entangled, etc.) described in Section 2.

The training uses a well-known cosine-annealing learning rate schedule and standard PyTorch Lightning callbacks for logging and checkpointing. Upon completion, the final checkpoints were saved for subsequent classification.

To evaluate the pretraining performace we introduce a metric called Hermitian distance, that measures the deviation of a matrix from being perfectly Hermitian by computing the average Frobenius norm of the difference between the matrix and its conjugate transpose:

$$h = \frac{1}{b} \sum_{k=1}^{b} \sqrt{||\boldsymbol{A}_k - \boldsymbol{A}_k^\dagger||_F}, \tag{7}$$

where b is the number of matrices in the batch, and $\mathbf{A}_k$ denotes the k-th complex matrix. The notation $\mathbf{A}_k^\dagger$ refers to its conjugate transpose, and $\|\cdot\|_F$ indicates the Frobenius norm. For a Hermitian matrix, $\mathbf{A}_k = \mathbf{A}_k^\dagger$, which makes the norm vanish. When the matrix is split into real and imaginary parts, $\mathbf{A}_k = \mathbf{R}_k + i\,\mathbf{I}_k$, Hermiticity requires $\mathbf{R}_k$ to be symmetric $\left(\mathbf{R}_k = \mathbf{R}_k^T\right)$ and $\mathbf{I}_k$ to be antisymmetric $\left(\mathbf{I}_k = -\mathbf{I}_k^T\right)$. Consequently, the quantity $\mathbf{A}_k - \mathbf{A}_k^\dagger$ captures deviations from these symmetries, and its Frobenius norm measures how far the matrix is from being perfectly Hermitian. Averaging over all matrices in the batch yields the final distance h.

We used this metric to evaluate our pretraining process by assessing how well the pretrained model preserves the Hermitian structure of the data when 15% of the matrix is reconstructed by a network. A lower Hermitian distance indicates that the intrinsic mathematical properties are maintained, serving as a meaningful indicator of the quality of pretraining.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

3.3 Classifier Training

After pretraining, we fine-tune the learned Transformer weights for a downstream binary classification task. The core steps are:

- Load the pre-trained Transformer weights into a new model that augments the Transformer encoder with a feed-forward classification head (two-class output).
- Use the datasets of 1.9 mln (for $\mathbb{C}^2 \otimes \mathbb{C}^2$ and $\mathbb{C}^4 \otimes \mathbb{C}^4$), 2 mln (for $\mathbb{C}^2 \otimes \mathbb{C}^3$), and 3.5 mln (for $\mathbb{C}^3 \otimes \mathbb{C}^3$), described in detail in Table 1 in the *Classification* subsection. This data is separate from the pretraining data.
- Split the dataset into training (90%), validation (5%), and test (5%) partitions.
- Train the network using cross-entropy loss and the same PyTorch Lightning setup, with logging, checkpointing, and cosine-annealing learning rate schedule.

This two-stage approach leverages the pre-trained Transformer's learned representation of the quantum matrices, enhancing the performance of the downstream classification task.

4 Results and discussion

To determine the final results of the pretraining and classification, we used the separate data subset that was not used during the training phase neither for training nor for validation and testing. This subset comprises of 100,000 samples for each of the classes. The final evaluation was carried out for both pretraining and classification after the entire training process was completed.

Pretraining During the pretraining phase, both the loss function and the Hermitian distance improved significantly. Table 2 shows the averaged results of the Hermitian distance metric for the pretraining phase. The results are consistent and show that, overall, the pretraining process ended up generating models that are able to reconstruct the Hermitian structure of the data. Surprisingly, the Hermitian distances achieved very low values during the early pretraining. We did not observe any significant improvement in this metric after the first few epochs, whereas the loss function continued to decrease. This suggests that the model learned the Hermitian structure of the data very quickly, and the loss function was optimized to a greater extent.

Classification The classification results are presented in Table 3. It is important to emphasize that, while results are presented for each type of state, the classification was deliberately kept binary–entangled vs. separable–as this is the relevant distinction for practical applications. The results show that the model was able to classify the states with very high accuracy. The results are consistent across all dimensions and classes. The only errors appear for the pure

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

Group name	$\mathbb{C}^2 \otimes \mathbb{C}^2$	$\mathbb{C}^2 \otimes \mathbb{C}^3$	$\mathbb{C}^3 \otimes \mathbb{C}^3$	$\mathbb{C}^4 \otimes \mathbb{C}^4$
Untrained				
sep	6.686	3.718	8.101	13.922
general-ent	6.704	3.716	8.098	13.921
werner-ent	6.645		8.131	13.908
max-ent	6.693		8.102	13.921
horodecki-bound			8.104	
horodecki-ent			8.102	
Pretrained				
sep	0.265	0.364	0.419	0.458
general-ent	0.189	0.187	0.167	0.130
werner-ent	0.183		0.222	0.248
max-ent	0.296		0.449	0.478
horodecki-bound			0.120	
horodecki-ent			0.122	

Table 2. Averaged Hermitian distances for untrained and fully pretrained models.

separable states where a few of states were misclassified as entangled, and for max-entanglement states in $\mathbb{C}^4 \otimes \mathbb{C}^4$ group where a few states were misclassified as separable. This confirms the validity of our approach, as the model effectively captures the structural properties of quantum states and generalizes well across different state classes and dimensions, with only minimal misclassification in specific cases.

Group name	$\mathbb{C}^2 \otimes \mathbb{C}^2$	$\mathbb{C}^2 \otimes \mathbb{C}^3$	$\mathbb{C}^3 \otimes \mathbb{C}^3$	$\mathbb{C}^4 \otimes \mathbb{C}^4$
sep	99.995%	99.998%	100%	99.941%
general-ent	100%	100%	100%	100%
werner-ent	100%		100%	100%
max-ent	100%		100%	99.851%
horodecki-bound			100%	
horodecki-ent			100%	

Table 3. Accuracy of binary classification (entangled vs separated) for each data group.

To further verify our results, we tested whether the deep fine-tuning during classification training provides an additional learning benefit beyond the pretraining phase. Specifically, we took a pretrained model and froze all its layers except for the final classification layer, which was then fine-tuned on labeled data. This approach aimed to determine whether the pretrained representations alone were sufficient for entanglement classification or if further adaptation was necessary. The fine-tuned model performed nearly perfectly on $\mathbb{C}^2 \otimes \mathbb{C}^2$ and $\mathbb{C}^2 \otimes \mathbb{C}^3$ states but struggled with $\mathbb{C}^3 \otimes \mathbb{C}^3$ states, achieving around 85% accuracy. A de-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

tailed breakdown revealed that while entangled states were consistently classified correctly, the model frequently misclassified separable states. This suggests that while the pretrained model captures general entanglement patterns, adapting deeper layers during training may be crucial for distinguishing subtle features in higher-dimensional separable states.

Discussion Our work is closely related to [8], which also explores automated entanglement classification. However, we achieve significantly better results, with near-perfect accuracy compared to their reported range of 62–88%. While their approach struggled with deep learning, we successfully adapted transformers into a highly effective classification method. An important distinction is that the presence of bound entangled states in our dataset does not degrade performance. Additionally, our dataset is substantially larger, containing millions of states, whereas theirs consisted of only 3,254. Their dataset generation method, in principle, allows for bound entangled states in any dimension, while our approach focuses on a specific family of states for the $\mathbb{C}^3 \otimes \mathbb{C}^3$ case. However, their generation method is much more computationally expensive. Overall, our results are consistent with theirs, but we extend the approach significantly, demonstrating the feasibility of deep learning for entanglement classification at a much larger scale.

Our work serves as an example of the successful integration of machine learning techniques with quantum information science. Similar approaches–whether in developing quantum information models [2] or solving quantum computing problems [22]–represent a promising direction for the field. We expect that this fusion of computational methods and quantum theory will gain increasing prominence, much like the impact of machine learning in computational chemistry.

Our approach addresses a different but related problem compared to [9]. Their work focuses on generating entanglement witnesses for specific types of quantum states and specially structured witnesses. While their method is well-executed, it is inherently limited, as reflected in their choice of states. In contrast, our approach is applied to a significantly larger dataset, allowing for a broader and more flexible classification of entanglement. However, their method extends to multipartite entanglement, an area we have not yet explored.

Non-ML classificarion To assess the difficulty of the problem and to benchmark the performance of proposed approach against classical, we additionally tested a non-Machine Learning method: Logistic Regression. It was selected as a simple, interpretable classifier based on a linear decision boundary in the feature space. Importantly, Logistic Regression does not capture higher-order correlations or complex feature interactions, providing a baseline for how much structure in the quantum state matrices can be detected via linear separation alone.

For the experiment the $\mathbb{C}^3 \otimes \mathbb{C}^3$ dataset was used. The classifier was trained on the same training dataset used in the basic deep learning setup. The model was optimized using cross-entropy loss with L2 regularization. No explicit feature engineering or non-linear kernelization was applied.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

The results, summarized in Table 4, reveal a significant drop in classification performance compared to the proposed approach. It can be seen, that while certain entangled states exhibit simple structures that can be identified linearly, the general problem of entanglement classification, especially distinguishing pure separable states and maximally entangled states, remains hard for simple linear models.

Group name	Accuracy (%)
sep	41.93%
general-ent	75.22%
horodecki-bound	100.00%
horodecki-ent	100.00%
werner-ent	100.00%
max-ent	53.98%

Table 4. Logistic Regression classification accuracy across different quantum state groups. While the model achieved perfect classification on some classes, its performance was substantially lower on others. Separable states achieved only around 41.9% accuracy, and maximally entangled states (max-ent) were classified with 53.9% accuracy. General entangled states showed moderate success, with 75.2% accuracy.

Overall, while some structured quantum states can be linearly distinguished, robust, high-accuracy classification across a diverse set of quantum states requires models with greater capacity.

Limitations While our Quantum-aware Transformer shows near-perfect accuracy on a large synthetic benchmark, several caveats warrant mention. (i) All experiments presume full state tomography, which is an idealisation. (ii) Training, validation, and test sets were produced by a synthetic pipeline and were limited to bipartite systems up to $\mathbb{C}^4 \otimes \mathbb{C}^4$; hence, generalisation to higher-dimensional or multipartite states and to independently synthesised data, remains unverified. (iii) Finally, the present study tackles only binary separable versus entangled discrimination and, apart from the fixed Horodecki family, does not distinguish free from bound entanglement or quantify entanglement strength. We regard these open issues as priorities for the next phase of our research.

Conclusions We have demonstrated that transformer-based neural networks can effectively classify bipartite quantum states as entangled or separable by learning directly from quantum state matrices. By leveraging a masked autoencoding pretraining strategy, our model captures the structural properties of density matrices, achieving near-perfect classification accuracy across various state types and dimensions. These results highlight the potential of modern deep learning architectures for quantum information processing, paving the way for scalable, data-driven approaches to entanglement detection and beyond.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

Acknowledgments. This project was supported by the National Science Center (NCN), Poland, under Projects: Sonata Bis 10, No. 2020/38/E/ST3/00269 (L.P.)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bennett, C.H., DiVincenzo, D.P., Mor, T., Shor, P.W., Smolin, J.A., Terhal, B.M.: Unextendible product bases and bound entanglement. Physical Review Letters **82**(26), 5385–5388 (1999). https://doi.org/10.1103/PhysRevLett.82.5385
2. Cholewa, M., Gawron, P., Głomb, P., Kurzyk, D.: Quantum hidden Markov models based on transition operation matrices. Quantum Information Processing **16**(4), 101 (Mar 2017). https://doi.org/10.1007/s11128-017-1544-8
3. Cramer, M., Plenio, M.B., Flammia, S.T., Somma, R., Gross, D., Bartlett, S.D., Landon-Cardinal, O., Poulin, D., Liu, Y.K.: Efficient quantum state tomography. Nature Communications **1**(1), 149 (2010). https://doi.org/10.1038/ncomms1147
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805
5. Gawron, P., Kurzyk, D., Pawela, Ł.: QuantumInformation.jl—a julia package for numerical computation in quantum information theory. PLOS ONE **13**(12), e0209358 (dec 2018). https://doi.org/10.1371/journal.pone.0209358
6. Gisin, N., Thew, R.: Quantum communication. Nature Photonics **1**, 165–171 (2007). https://doi.org/10.1038/nphoton.2007.22
7. Goel, R., Xiao, Y., Ramezani, R.: Transformer models as an efficient replacement for statistical test suites to evaluate the quality of random numbers. In: 2024 International Symposium on Networks, Computers and Communications (ISNCC). pp. 1–6 (2024). https://doi.org/10.1109/ISNCC62547.2024.10758985
8. Goes, C.B.D., Canabarro, A., Duzzioni, E.I., Maciel, T.O.: Automated machine learning can classify bound entangled states with tomograms. Quantum Information Processing **20**(3), 99 (Mar 2021). https://doi.org/10.1007/s11128-021-03037-9
9. Greenwood, A.C., Wu, L.T., Zhu, E.Y., Kirby, B.T., Qian, L.: Machine-learning-derived entanglement witnesses. Physical Review Applied **19**(3), 034058 (2023). https://doi.org/10.1103/PhysRevApplied.19.034058
10. Gühne, O., Tóth, G.: Entanglement detection. Physics Reports **474**(1-6), 1–75 (2009). https://doi.org/10.1016/j.physrep.2009.02.004
11. Horodecki, M., Horodecki, P., Horodecki, R.: Separability of mixed states: necessary and sufficient conditions. Physics Letters A **223**(1-2), 1–8 (1996). https://doi.org/10.1016/S0375-9601(96)00706-2
12. Horodecki, P., Horodecki, M., Horodecki, R.: Mixed-state entanglement and distillation: is there a "bound" entanglement in nature? Physical Review Letters **80**(24), 5239–5242 (1998). https://doi.org/10.1103/PhysRevLett.80.5239
13. Horodecki, P., Horodecki, M., Horodecki, R.: Bound entanglement can be activated. Physical Review Letters **82**(5), 1056 (1999). https://doi.org/10.1103/PhysRevLett.82.1056
14. Horodecki, R., Horodecki, P., Horodecki, M., Horodecki, K.: Quantum entanglement. Reviews of Modern Physics **81**(2), 865–942 (2009). https://doi.org/10.1103/RevModPhys.81.865

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15

15. Kukulski, R., Nechita, I., Pawela, Ł., Puchała, Z., Życzkowski, K.: Generating random quantum channels. Journal of Mathematical Physics **62**(6) (2021). https://doi.org/10.1063/5.0038838
16. Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information. Cambridge University Press, 10th anniversary ed. edn. (2010)
17. Nielsen, M.A., Chuang, I.L.: Quantum computation and quantum information. Cambridge university press (2010)
18. Ozols, M.: How to generate a random unitary matrix. http://home.lu.lv/sd20008 (2009), accessed: 2025-02-12
19. Peres, A.: Separability criterion for density matrices. Physical Review Letters **77**, 1413–1415 (1996). https://doi.org/10.1103/PhysRevLett.77.1413
20. Plenio, M.B., Virmani, S.: An introduction to entanglement measures. Quantum Information & Computation **7**(1-2), 1–51 (2007). https://doi.org/10.5555/2011706.2011707
21. Puchała, Z., Jenčová, A., Sedlák, M., Ziman, M.: Exploring boundaries of quantum convex structures: Special role of unitary processes. Physical Review A **92**, 012304 (2015). https://doi.org/10.1103/PhysRevA.92.012304
22. Śmierzchalski, T., Pawela, Ł., Puchała, Z., Trzciński, T., Gardas, B.: Post-error correction for quantum annealing processor using reinforcement learning. In: Groen, D., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds.) Computational Science – ICCS 2022. pp. 261–268. Springer International Publishing (2022)
23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
24. Vidal, G., Werner, R.F.: A computable measure of entanglement. Physical Review A **65**(3), 032314 (2002). https://doi.org/10.1103/PhysRevA.65.032314
25. Watrous, J.: The theory of quantum information. Cambridge university press (2018)
26. Życzkowski, K.: Volume of the set of separable states. ii. Physical Review A **60**(5), 3496 (1999). https://doi.org/10.1103/PhysRevA.60.3496
27. Zyczkowski, K., Horodecki, P., Sanpera, A., Lewenstein, M.: On the volume of the set of mixed entangled states. Physical Review A **58**(arXiv: quant-ph/9804024), 883 (1998). https://doi.org/10.1103/PhysRevA.58.883
28. Życzkowski, K., Horodecki, P., Sanpera, A., Lewenstein, M.: Volume of the set of separable states. Physical Review A **58**(2), 883 (1998). https://doi.org/10.1103/PhysRevA.58.883

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97570-7_15