

Article

Self-Attention over Parallel Dense Embeddings for High-Dimensional Omic Data

Kamal Elatifi ¹, Nicolas Jäger Gallego ², Alex Sánchez-Pla ² and Ferran Reverter ^{2,*} ¹ Biostatistics and Bioinformatics, Open University of Catalonia, 08018 Barcelona, Spain; kelatifi0@uoc.edu² Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain; nicolasjaergergallego@gmail.com (N.J.G.); asanchez@ub.edu (A.S.-P.)* Correspondence: freverter@ub.edu

Abstract

High-dimensional omic datasets present major challenges for machine learning due to their sparse biological signal, strong feature heterogeneity, and high dimensionality. In this work, we propose PLAT (Parallel Latent Attention Transformer), a neural architecture for high-dimensional tabular transcriptomic data. The model projects input gene expression features into multiple parallel latent representations, each processed independently through self-attention to capture complementary feature interactions while maintaining moderate model complexity. The proposed architecture was evaluated using both controlled Negative Binomial simulations designed to reproduce RNA-seq overdispersion and the TCGA-BRCA breast cancer dataset comprising 499 patients and 4376 gene expression variables for ER+/ER− classification. Comparative analyses against a baseline multilayer perceptron and a lightweight FT-Transformer showed that PLAT achieves competitive predictive performance while maintaining a comparable number of trainable parameters. Simulation experiments further indicate that its main advantage is concentrated in specific high-dimensional settings with an intermediate proportion of informative features. To assess model interpretability, we additionally performed a SHAP-based analysis of the baseline MLP and compared it with the attention-derived gene rankings. Although both models identified largely different sets of predictive genes, functional enrichment analyses consistently highlighted biological processes and disease pathways associated with breast cancer, supporting the biological relevance of the learned latent representations. These results suggest that PLAT provides an effective and interpretable framework for high-dimensional transcriptomic classification.

Keywords: tabular data; transformers; self-attention; transcriptomics

Academic Editor: Douglas O'Shaughnessy

Received: 15 June 2026

Revised: 5 August 2026

Accepted: 6 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

One of the primary challenges associated with omics data lies in its pronounced heterogeneity. Biological information can be represented in diverse formats, including sequences, graphs, and matrices—commonly referred to as tabular data within the deep learning framework [1,2].

In this work, we explore an adaptation of Transformer-style self-attention architectures for tabular data. While Transformers [3,4] have been predominantly applied to sequential and image data [5–7], the mechanism of self-attention—modeling interactions between elements of an input sequence—can also be leveraged to capture dependencies between features in structured tabular datasets.

Various approaches addressing this problem have already been proposed [8–10], including models such as the FT Transformer [11] and the TabTransformer [12]. In essence, these methods learn dedicated embeddings for each feature in tabular data. We introduce an alternative strategy for learning data embeddings that employs multiple parallel hidden layers, each directly connected to the input layer and designed with a limited number of nodes to capture complementary feature representations. The proposed architecture is motivated by the hypothesis that explicitly incorporating attention mechanisms over latent representations derived from high-dimensional tabular data can enhance the model’s ability to capture complex interactions among features compared to traditional dense architectures. In this context, the design prioritizes parameter efficiency relative to large fully connected models, enables the modeling of relationships without representing each feature as an explicit token, and is particularly well-suited for extremely high-dimensional settings such as omics datasets.

2. Materials and Methods

2.1. Self-Attention over Parallel Dense Embeddings

The overall architecture of the proposed method consists of the following steps:

1. **Input Layer:** Each input sample consists of p features (columns). The input layer directly receives these features.
2. **Parallel Dense Layers:** The input layer is mapped to q parallel dense embedding branches of dimension d . These hyperparameters are chosen to provide a compact latent representation, thereby reducing the computational burden and parameter count relative to a standard fully connected network. These layers transform the input into multiple representations, producing q vectors of dimension d . Formally:

$$h_i = \text{ReLU}(W_i x + b_i), \quad i = 1, \dots, q,$$

where ReLU is the Rectified Linear Unit activation function used to introduce nonlinearity, W_i denotes the weight matrix of the i -th parallel branch, x the input vector, and b_i the bias vector of the i -th parallel branch.

3. **Reshaping into Sequence (Embedding Matrix):** The outputs of the parallel layers are stacked to form a matrix H of shape $q \times d$, which can be interpreted as a sequence of q “tokens” with embedding dimension d . This matrix acts as the input embedding for the attention mechanism.
4. **Multi-head Attention Layer:** We apply Transformer-style self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V.$$

Let $H = [h_1, \dots, h_q]^T$ denote the matrix whose rows are the embeddings produced by the q parallel branches. These embeddings are treated as the input tokens of the multi-head self-attention layer and are projected into query, key, and value representations using the learnable projection matrices W_i^Q , W_i^K , and W_i^V :

$$\text{head}_i = \text{Attention}(HW_i^Q, HW_i^K, HW_i^V), \quad i = 1, \dots, h,$$

where W_i^Q , W_i^K , and W_i^V are the learnable query, key, and value projection matrices for the i -th attention head and h denotes the number of attention heads.

The outputs of the h attention heads are concatenated and projected through the learnable output projection matrix W^O :

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O,$$

where $\text{Concat}(\cdot)$ concatenates the outputs of all attention heads and W^O denotes the learnable output projection matrix. This allows the model to capture interactions between the parallel dense representations.

5. Residual Connections and Normalization: We employ residual connections and layer normalization after the attention layer, as in standard Transformers [3].
6. Feed-Forward Network (FFN): Following the attention layer, each token passes through a small feed-forward network (two dense layers with a ReLU activation in between), again with residual connections and layer normalization [3].
7. Output Layer: The resulting sequence is flattened into a single vector, which is then fed into a final dense layer with sigmoid activation for binary classification (or softmax for multi-class tasks).

2.2. Architectures

For the purpose of this study, three models are compared with each other: an MLP baseline, a lightweight FT-Transformer [11] and our proposed parallel embedding architecture, which will be called PLAT (Parallel Latent Attention Transformer) from now on.

The first model corresponds to a baseline Multi-Layer Perceptron (MLP), implemented as a fully connected feed-forward neural network [13]. The architecture consists of an input linear transformation from the feature space of dimension p to q hidden neurons, followed by a ReLU activation function and a final layer producing the output prediction. This architecture provides a simple reference model with relatively low computational complexity.

The second architecture corresponds to a lightweight latent-token variant of FT-Transformer inspired by the original architecture. Each input feature is first embedded independently into a d -dimensional representation. Unlike the original FT-Transformer, the resulting feature embeddings are subsequently projected into a reduced latent token space before the self-attention layer. This modification was introduced to make the architecture computationally feasible for high-dimensional transcriptomic data while maintaining a comparable number of trainable parameters to PLAT. The projection dimension was selected according to the dimensionality of each dataset so that the resulting FT-Transformer had a number of trainable parameters comparable to that of the corresponding PLAT model, ensuring a fair architectural comparison. Specifically, in the simulation study, different projection dimensions were adopted for the $p = 1000$ and $p = 5000$ scenarios, whereas for the experimental RNA-seq dataset ($p = 4376$ genes), the feature sequence was projected from 4376 to 256 latent tokens. Consequently, self-attention was computed over the compressed latent representation rather than over the original feature set, substantially reducing computational and memory requirements. For the real dataset, this reduced the attention matrices from 4376×4376 (approximately 19.1 million elements) to 256×256 (65536 elements), while maintaining the ability to capture long-range feature interactions. The resulting contextualized representation was globally pooled and fed to a final dense layer with a sigmoid activation for binary classification.

The third architecture corresponds to the proposed tabular attention model with parallel dense embedding layers, PLAT, which incorporates an attention mechanism specifically designed for tabular data (see Figure 1). The model first projects the p input features into q parallel fully connected layers, each with a lower-dimensional latent representation d . The resulting embeddings are fed into a Multi-Head Attention layer with h attention heads. Each head can be interpreted as an independent replica of the same attention mechanism

operating on different learned projections of the latent space. This design enables the model to learn multiple complementary feature interactions in parallel. The outputs generated by the different heads are then concatenated and projected into a unified representation. After the attention stage, the latent representation is refined using a feed-forward network composed of two hidden transformations.

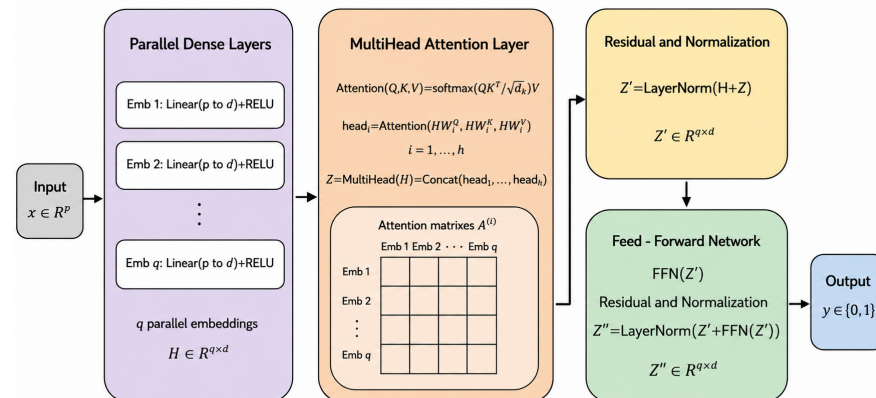


Figure 1. Overview of the proposed architecture. Input features are projected through multiple parallel fully connected layers with a reduced number of nodes, reshaped into structured latent representations, and processed by a self-attention module via query (Q), key (K), and value (V) projections to produce context-aware embeddings. These embeddings can then be further processed by additional layers, such as attention-based or dense layers, to complete the architecture.

Finally, the extracted features are mapped to the prediction output through a fully connected layer.

2.3. Simulation Scenarios

We propose a generative framework to simulate RNA-seq count data with controllable two-class classification difficulty. The model combines a latent variable formulation for class assignment with a gene-wise Negative Binomial model for count generation, capturing overdispersion and heterogeneity typical of transcriptomic data.

2.3.1. Latent Variable and Class Assignment

Let n denote the number of samples. For each sample $i = 1, \dots, n$, we introduce a latent variable u_i , assumed to follow a standard normal distribution:

$$u_i \sim \mathcal{N}(0, 1),$$

representing an unobserved biological gradient.

Binary class labels $y_i \in \{0, 1\}$ are generated by thresholding the latent variable after adding Gaussian noise:

$$y_i = \mathbf{1}(u_i + \epsilon_i > 0), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\mathbf{1}(\cdot)$ is the indicator function. The variance σ^2 determines the amount of label noise and hence the degree of overlap between the two classes, with larger values leading to a more difficult classification task.

2.3.2. Gene Expression Model

We assume that a fraction $r \in (0, 1)$ of genes are informative with respect to the class label. Let $\mathcal{S} \subset \{1, \dots, p\}$ denote the set of informative genes with $|\mathcal{S}| = \lfloor rp \rfloor$.

For each gene $j = 1, \dots, p$, we define:

- a continuous effect $A_j \sim \mathcal{N}(0, \sigma_A^2)$,
- a class-specific effect

$$B_j = \begin{cases} \tilde{B}_j, & j \in \mathcal{S}, \\ 0, & j \notin \mathcal{S}, \end{cases}$$

where $\tilde{B}_j \sim \mathcal{N}(\mu_B, \sigma_B^2)$.

The log-mean expression for gene j in sample i is given by

$$\log \mu_{ij} = u_i A_j + y_i B_j.$$

2.3.3. Negative Binomial Observation Model

Observed gene expression counts X_{ij} are generated from a Negative Binomial distribution:

$$X_{ij} \sim \text{NB}(\mu_{ij}, \theta_j),$$

where $\mu_{ij} = \exp(u_i A_j + y_i B_j)$ is the mean parameter and θ_j is a gene-specific dispersion parameter. We assume

$$\theta_j \sim \text{Gamma}(\alpha_\theta, \beta_\theta),$$

allowing for heterogeneous overdispersion across genes [14]. Under this parametrization, the variance satisfies

$$\text{Var}(X_{ij}) = \mu_{ij} + \frac{\mu_{ij}^2}{\theta_j}.$$

2.3.4. Calibration of Classification Difficulty

This calibration step ensures that the simulated datasets exhibit controlled separability, avoiding trivial classification scenarios and enabling systematic benchmarking across varying difficulty regimes.

To control the classification difficulty, we calibrate the noise parameter σ using a bisection procedure. Starting from an initial value of $\sigma = 0.3$, a dataset is generated and split into training and test subsets. A logistic regression classifier is trained on the log-transformed data and evaluated on the test set to compute the empirical classification error

$$\hat{\epsilon}(\sigma) = 1 - \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} \mathbf{1}(\hat{y}_i = y_i).$$

The parameter σ is iteratively updated until

$$|\hat{\epsilon}(\sigma) - \epsilon^*| < \delta,$$

where ϵ^* is the target error rate ($\epsilon^* = 0.25$) and δ is a user-defined tolerance controlling the convergence criterion ($\delta = 0.05$).

2.3.5. Hyperparameter Selection and Ablation Study

To further validate the PLAT architecture, we performed a comprehensive hyperparameter study considering the number of parallel branches ($q \in 1, 2, 4, 8$), embedding dimensions ($d \in 8, 16, 32, 64$), and attention heads ($h \in 1, 4$), resulting in 32 architectural configurations evaluated over 10 random seeds in two simulation scenarios ($p = 1000$ and $p = 5000$). Hyperparameter selection was conducted exclusively on the training data using an internal five-fold cross-validation procedure, ensuring that the test set remained completely independent throughout model selection (see Algorithm 1). The results consistently

demonstrated the effectiveness of the proposed parallel architecture, with predictive performance increasing monotonically as the number of parallel branches increased (Figure 2). In the $p = 1000$ scenario, the mean AUC improved from 0.791 for a single branch ($q = 1$) to 0.849 for four branches ($q = 4$) and the improvement was statistically significant (paired Wilcoxon signed-rank test, $p = 0.002$). Since $q = 1$ corresponds to the architecture without parallel branches, these results provide a direct ablation study supporting the contribution of the proposed parallel design (Figure 2). Furthermore, no meaningful performance differences were observed between one and four attention heads, indicating that a single attention head is sufficient to achieve competitive predictive performance while substantially reducing the number of trainable parameters and overall model complexity.

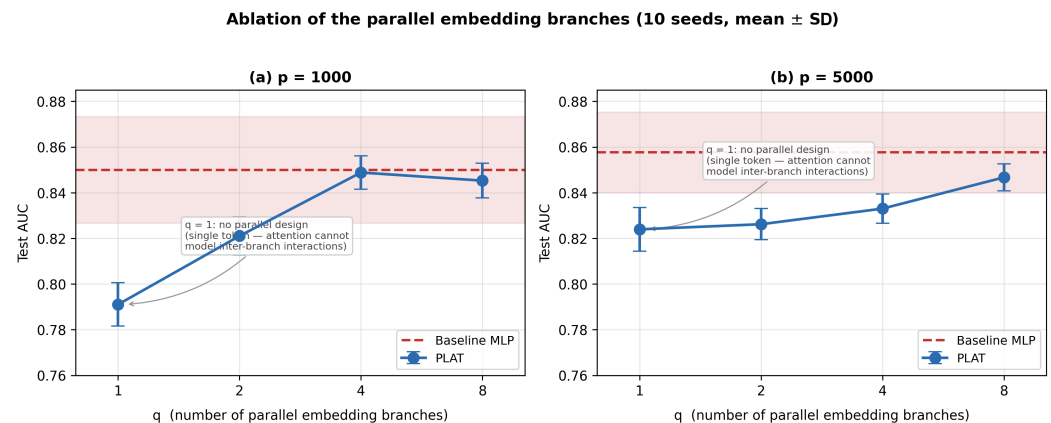


Figure 2. Ablation study of the proposed PLAT architecture by varying the number of parallel embedding branches q . Mean test AUC ($\pm$ standard error) over 10 random seeds is shown for two simulation settings: (a) $p = 1000$ and (b) $p = 5000$. The case $q = 1$ corresponds to a single embedding branch without the proposed parallel design, providing a direct ablation of the architecture. Increasing the number of parallel branches consistently improves predictive performance, demonstrating the contribution of the parallel latent representations. The dashed horizontal line indicates the performance of the baseline MLP, and the shaded band represents ± 1 standard deviation.

Algorithm 1 Nested hyperparameter selection and evaluation protocol

```

1: for each simulation scenario ( $p = 1000, 5000$ ) do
2:   for each random seed ( $1, \dots, 10$ ) do
3:     Generate simulated dataset
4:     Split into training and independent test sets
5:     for all  $(q, d, h) \in \{1, 2, 4, 8\} \times \{8, 16, 32, 64\} \times \{1, 4\}$  do
6:       Estimate mean AUC by 5-fold cross-validation on the training set
7:     end for
8:     Select the configuration maximizing the mean cross-validation AUC
9:     Retrain PLAT on the complete training set using the selected configuration
10:    Evaluate once on the untouched test set
11:  end for
12:  Report mean  $\pm$  standard deviation of the test AUC over the 10 seeds
13: end for

```

Table 1 reports the best-performing PLAT hyperparameter configuration identified for each simulation scenario. These configurations were selected based on the average predictive performance obtained during the hyperparameter search and were subsequently used in all comparative simulation experiments. The complete results for all evaluated hyperparameter combinations are available in Supplementary Excel File S1.

Table 1. Best PLAT hyperparameter configuration selected for each simulated dimensionality.

Scenario (p)	Best q	Best d	Best h	Parameters
1000	4	64	4	289985
5000	8	16	1	642481

Interestingly, the optimal architecture depended on the dimensionality of the simulation scenario. While the $p = 1000$ setting favored a larger embedding dimension and multiple attention heads, the $p = 5000$ scenario selected a considerably more compact configuration. This suggests that, as feature dimensionality increases, simpler attention mechanisms may provide a better trade-off between representational capacity and model complexity.

2.4. Real Data

To validate the predictive capability of the proposed architecture in a real biological setting, the model was evaluated using RNA-seq gene expression data obtained from the TCGA-BRCA cohort [15] through the Genomic Data Commons (GDC) portal [16]. After preprocessing and filtering, the resulting dataset comprised 499 breast cancer patients and 4376 normalized gene expression features. The target variable corresponded to Estrogen Receptor (ER) status, including 386 ER-positive samples (77.4%) and 113 ER-negative samples (22.6%).

The two classes represent clinically distinct breast cancer subtypes [17,18]. ER+ (Estrogen Receptor positive) tumors express estrogen receptors on the surface of cancer cells and are generally associated with a better prognosis and responsiveness to hormone-based therapies such as tamoxifen or aromatase inhibitors. In contrast, ER− (Estrogen Receptor negative) tumors do not express these receptors and are often associated with more aggressive disease progression and fewer targeted treatment options, commonly requiring chemotherapy-based approaches. Consequently, accurate discrimination between these two groups is highly relevant from both a diagnostic and therapeutic perspective.

3. Results

3.1. Simulation Results

Table 2 summarizes the predictive performance of the three models under the two simulated scenarios. For PLAT, the reported results correspond to the best hyperparameter configuration identified independently for each scenario through the internal hyperparameter selection procedure described in Section 2.3.5. In the moderate-dimensional setting ($p = 1000$), both attention-based models outperformed the baseline MLP in terms of ROC-AUC and PR-AUC. PLAT achieved the highest ROC-AUC (0.887) and the best balanced accuracy (0.821), while FT-Transformer obtained the highest PR-AUC (0.887) and sensitivity (0.869), at the expense of a lower specificity. In the high-dimensional setting ($p = 5000$), the performance differences became smaller. FT-Transformer achieved the highest ROC-AUC (0.878), PR-AUC (0.890), and specificity (0.759), whereas the baseline MLP obtained the highest balanced accuracy (0.787). The three architectures exhibited comparable predictive performance in the more challenging high-dimensional scenario, with no single model consistently outperforming the others across all evaluation metrics.

Pairwise Wilcoxon signed-rank tests (Table 3) performed across the 10 random seeds revealed no statistically significant differences between any pair of models in either simulation scenario ($p > 0.05$). Although small differences in average ROC-AUC were observed—favoring PLAT in the $p = 1000$ scenario and FT-Transformer in the $p = 5000$ scenario—these differences were not statistically significant. Therefore, the results sug-

gest that the three architectures provide comparable predictive performance under the simulated conditions.

Table 2. Comparison of Baseline MLP, PLAT, and FT-Transformer under two simulated scenarios. Results are reported as mean $\pm$ standard deviation across repetitions.

Scenario	Model	Params	ROC-AUC	PR-AUC	Balanced Acc	Sensitivity	Specificity
$p = 1000$	Baseline MLP	128257	0.863 ± 0.074	0.863 ± 0.095	0.778 ± 0.095	0.805 ± 0.187	0.752 ± 0.131
	PLAT	289985	0.887 ± 0.071	0.885 ± 0.099	0.821 ± 0.047	0.853 ± 0.103	0.790 ± 0.115
	FT-Transformer	266081	0.883 ± 0.059	0.887 ± 0.072	0.783 ± 0.114	0.869 ± 0.086	0.698 ± 0.235
$p = 5000$	Baseline MLP	640257	0.864 ± 0.064	0.874 ± 0.054	0.787 ± 0.043	0.840 ± 0.075	0.734 ± 0.115
	PLAT	642481	0.859 ± 0.076	0.879 ± 0.060	0.766 ± 0.081	0.841 ± 0.059	0.690 ± 0.163
	FT-Transformer	649953	0.878 ± 0.056	0.890 ± 0.060	0.781 ± 0.060	0.804 ± 0.126	0.759 ± 0.184

Table 3. Pairwise Wilcoxon signed-rank tests comparing the ROC-AUC obtained by the three architectures across the 10 random seeds.

Scenario	Model A	Model B	Δ ROC-AUC (A – B)	p -Value
$p = 1000$	PLAT	Baseline MLP	0.0239	0.2422
	PLAT	FT-Transformer	0.0041	0.8262
	FT-Transformer	Baseline MLP	0.0198	0.2812
$p = 5000$	PLAT	Baseline MLP	−0.0049	0.6953
	PLAT	FT-Transformer	−0.0193	0.3613
	FT-Transformer	Baseline MLP	0.0145	0.5508

Further details of the comparative experiments are provided in Supplementary Excel File S2.

Following the comparative evaluation presented above, we further investigate two practical aspects of the proposed PLAT architecture. First, we assess how the proportion of informative variables influences predictive performance in high-dimensional settings, thereby evaluating the robustness of the attention mechanism to increasing levels of irrelevant features. Second, we analyze how lightweight attention configurations affect the trade-off between model complexity and predictive performance. The objective of the analyses that follow is to contrast attention-based and non-attention architectures, rather than to perform an exhaustive comparison among different attention mechanisms. Accordingly, we focus on PLAT and the baseline MLP, which enables the effect of the proposed attention mechanism to be isolated relative to a standard feed-forward architecture.

3.1.1. Relevant Features Rate

Figure 3 illustrates the predictive performance of the baseline MLP and the proposed PLAT under different proportions of relevant features (5%, 10%, and 20%) using simulated datasets with $n = 100$ samples. The experiments evaluate how classification accuracy changes as the total number of features p increases, thereby progressively introducing a larger number of non-informative variables into the dataset.

Both models achieve their highest performance at intermediate dimensionalities, whereas predictive accuracy decreases as the number of irrelevant features becomes very large. This behavior is expected in high-dimensional settings, where noise and redundant variables can hinder the extraction of discriminative patterns. The shaded regions represent ± 1 standard deviation across repetitions. A notable pattern is observed in the scenario where 10% of the features are informative. In this setting, PLAT achieves, on average, higher predictive accuracy than the baseline MLP when the dimensionality reaches approximately 1000 features or more. As the total number of features continues to increase, however, this advantage gradually diminishes, and the performance of both models becomes increasingly similar. These results suggest that the benefit of the proposed attention mechanism is

concentrated within a specific operating regime characterized by a moderate signal-to-noise ratio and high dimensionality, rather than representing a consistent improvement across all simulation settings.

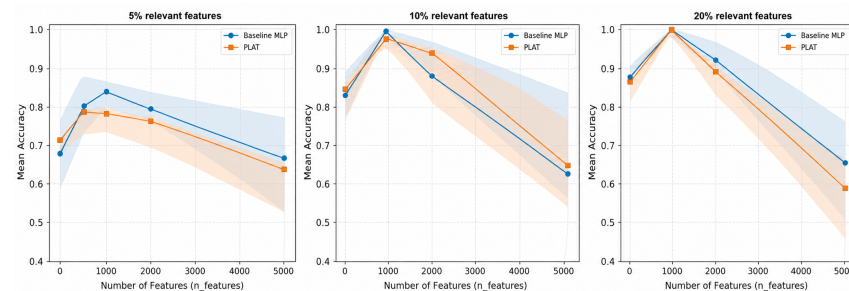


Figure 3. Comparison between the baseline MLP and the proposed attention model under different proportions of relevant features (5%, 10%, and 20%) using simulated data. Performance is evaluated as a function of the total number of features p . The shaded band represents ± 1 standard deviation.

The simulation results indicate that PLAT can provide an advantage under specific high-dimensional conditions, particularly when informative variables constitute an intermediate proportion of the feature space. However, this advantage is localized to a particular simulation regime and should not be interpreted as a general superiority over the baseline MLP across all considered scenarios.

3.1.2. Lightweight Architectures

An important question explored in this work is whether the proposed PLAT architecture can operate with fewer parameters than the baseline model while preserving competitive predictive performance.

The parameter counts reported in Table 4 were computed directly from four variants of the proposed PLAT architecture for the simulated Negative Binomial scenario ($n = 100$, $p = 1000$). All variants share the same embedding, attention, and classification modules while progressively reducing the size of the latent representation, resulting in models ranging from 12% to 104% of the baseline MLP in terms of trainable parameters. This controlled design enables a fair comparison of the trade-off between model complexity and predictive performance across different PLAT configurations.

Table 4. Comparison of lightweight attention-based architectures with respect to the baseline MLP model.

Configuration	Parameters (Thousands)	% of Baseline
Baseline MLP	128	100%
Very Lightweight Attention	15	12%
Lightweight Attention	30	23%
Semi-Lightweight Attention	64	50%
Standard Attention	133	104%

Figure 4 compares four attention-based architectures of increasing complexity under the Negative Binomial simulation setting ($n = 100$, $p = 1000$). Each architecture was trained for 50 epochs and evaluated over five independent random seeds. For each configuration, the total number of trainable parameters, mean classification accuracy, and mean AUC were computed, with error bars indicating one standard deviation across runs. The baseline MLP is shown for reference, allowing the comparison of predictive performance relative to model complexity.

The bars correspond to lightweight attention configurations containing substantially fewer parameters than the baseline architecture. As expected, reducing the number of parameters decreases the representational capacity of the model; however, some lightweight variants are still able to maintain competitive predictive performance. In particular, the medium-attention configuration achieves a favorable balance between complexity reduction and classification quality, preserving high AUC values while using considerably fewer parameters than the standard attention model.

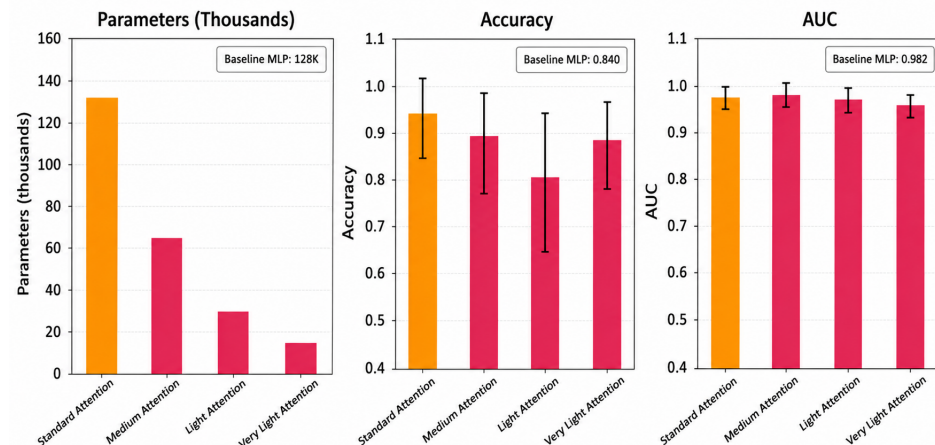


Figure 4. Comparison of PLAT and three lightweight attention variants with different model complexities. The figure reports the number of trainable parameters, classification accuracy, and AUC, with the baseline MLP included for reference.

The very lightweight configurations exhibit a larger reduction in parameter count but also show increased variability and slightly reduced predictive performance, suggesting that excessive compression may limit the ability of the model to learn robust latent representations from overdispersed data. In contrast, architectures with intermediate complexity appear to provide a more effective trade-off between computational efficiency and discriminative capability.

3.2. Real Data Results

In this section, we first present the quantitative performance metrics obtained from the comparative evaluation of the proposed PLAT architecture against the baseline MLP and the lightweight FT-Transformer. Although predictive accuracy and classification performance provide an essential benchmark for model evaluation, the most distinctive characteristic of the attention-based architecture lies in its enhanced interpretability. In particular, the attention mechanism offers additional insight into the relative importance and interactions of input variables, thereby improving the explainability of the model predictions [19,20]. Explainability via feature-wise attention is especially relevant in high-dimensional biological datasets, where understanding the contribution of specific features can be as important as achieving strong predictive performance.

3.2.1. Comparative Performance

Table 5 summarizes the predictive performance of the three models on the TCGA Breast Cancer dataset. All approaches achieved competitive results across the evaluated metrics. The baseline MLP obtained the highest values for Accuracy, Balanced Accuracy, Sensitivity, Specificity, and AUC, whereas PLAT achieved the highest PR-AUC. The complete performance metrics for the five independent stratified train/test splits are provided in Supplementary Excel File S3 to facilitate reproducibility. Importantly, the comparison

was conducted under equivalent model capacity, as the MLP, PLAT, and FT-Transformer contain 560385, 568929, and 571233 trainable parameters, respectively. Therefore, the observed differences are attributable to the architectural design rather than to disparities in the number of trainable parameters.

Table 5. Comparison of the predictive performance of the proposed PLAT model against the baseline FT-Transformer and MLP on the TCGA Breast Cancer dataset. Results are reported as mean $\pm$ standard deviation over five independent runs.

Metrics	PLAT	FT-Transformer	MLP
Accuracy	0.933 $\pm$ 0.015	0.918 $\pm$ 0.010	0.947 $\pm$ 0.017
Balanced Accuracy	0.900 $\pm$ 0.028	0.874 $\pm$ 0.024	0.920 $\pm$ 0.033
Sensitivity	0.959 $\pm$ 0.010	0.955 $\pm$ 0.012	0.967 $\pm$ 0.009
Specificity	0.842 $\pm$ 0.056	0.793 $\pm$ 0.053	0.873 $\pm$ 0.063
AUC	0.949 $\pm$ 0.027	0.936 $\pm$ 0.021	0.948 $\pm$ 0.030
PR-AUC	0.980 $\pm$ 0.013	0.973 $\pm$ 0.014	0.977 $\pm$ 0.016

3.2.2. Attention-Based Model Explainability

Figure 5 presents the average attention received by each parallel embedding branch of the proposed PLAT, separated according to ER+ and ER− samples. The analysis provides insight into how the self-attention mechanism distributes importance across the different latent embedding representations learned by the model.

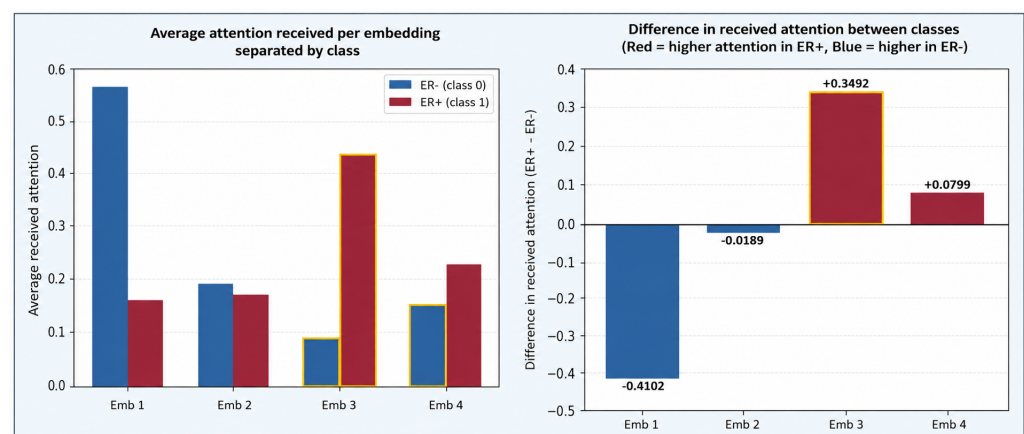


Figure 5. Average attention received by each embedding for ER+ and ER− samples (**left**) and the difference in attention between classes (**right**). The results show that Embedding 1 receives substantially higher attention in ER− samples, whereas Embedding 3 is strongly associated with ER+ samples, suggesting that different latent representations capture class-specific transcriptomic patterns.

The left panel shows the mean attention received by each embedding branch for both classes. A clear class-dependent attention pattern can be observed. Embedding 1 receives substantially higher attention in ER− samples, whereas Embedding 3 receives considerably greater attention in ER+ tumors. Embedding 4 also shows moderately increased attention for ER+ samples, while Embedding 2 exhibits relatively balanced attention between both classes. These results suggest that different embedding branches specialize in capturing transcriptomic patterns associated with distinct molecular subtypes.

The right panel displays the difference in received attention between classes, computed as ER+ minus ER−. Positive values indicate embeddings receiving more attention in ER+ samples, whereas negative values correspond to stronger attention in ER− tumors. The most pronounced difference is observed for Embedding 1, which is strongly associated with ER− samples, and Embedding 3, which shows the highest positive association with ER+

tumors. This complementary behavior supports the hypothesis that the parallel embedding strategy not only promotes representation diversity within the latent space, but also yields latent representations that more effectively discriminate between the two classes.

Taken together, the figure provides interpretability evidence for the proposed architecture, demonstrating that the attention mechanism captures biologically meaningful differences between ER+ and ER− breast cancer samples while distributing information across complementary latent embedding branches.

The activation analysis shown in Figures 6 and 7 provides additional interpretability regarding the internal behavior of the proposed PLAT on the TCGA Breast Cancer dataset.

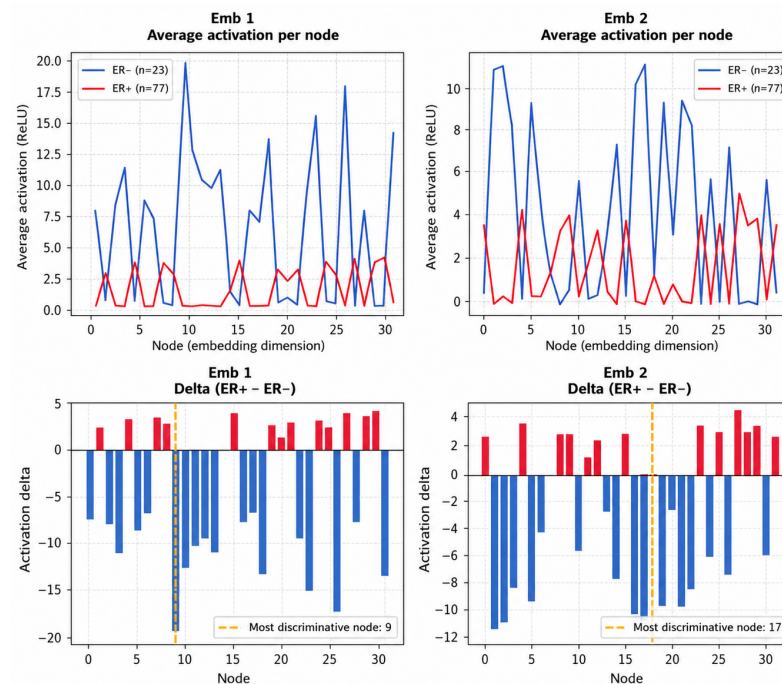


Figure 6. Average node activations for ER+ and ER− samples across the first two parallel embedding branches of PLAT (top row), together with the activation differences between classes (bottom row). Positive delta values indicate stronger activation in ER+ samples, whereas negative values indicate higher activation in ER− samples. The highlighted nodes correspond to the most discriminative embedding dimensions.

To analyze the behavior of the parallel embeddings, the test set was first divided according to the target class (ER− and ER+). For each parallel embedding branch, samples from each class were propagated independently through the corresponding embedding layer, and the mean activation of every embedding dimension (node) was computed across samples. The difference in mean activation between classes ($\text{delta} = \text{ER+} - \text{ER-}$) was then calculated for each node. Finally, the results were visualized by plotting the average activation profiles of both classes together with the node-wise activation differences, highlighting the embedding dimensions showing the strongest class-specific responses.

The top row presents the mean ReLU activations for each node separately for ER+ and ER− patients. Across the four embedding branches (Figures 6 and 7), clear differences can be observed between the two classes, indicating that the parallel architecture learns distinct latent representations associated with specific transcriptomic patterns. Some branches exhibit substantially stronger activations for ER− samples, whereas others respond preferentially to ER+ tumors. This complementary behavior suggests that biologically relevant information is distributed across multiple latent representations rather than concentrated within a single embedding branch.

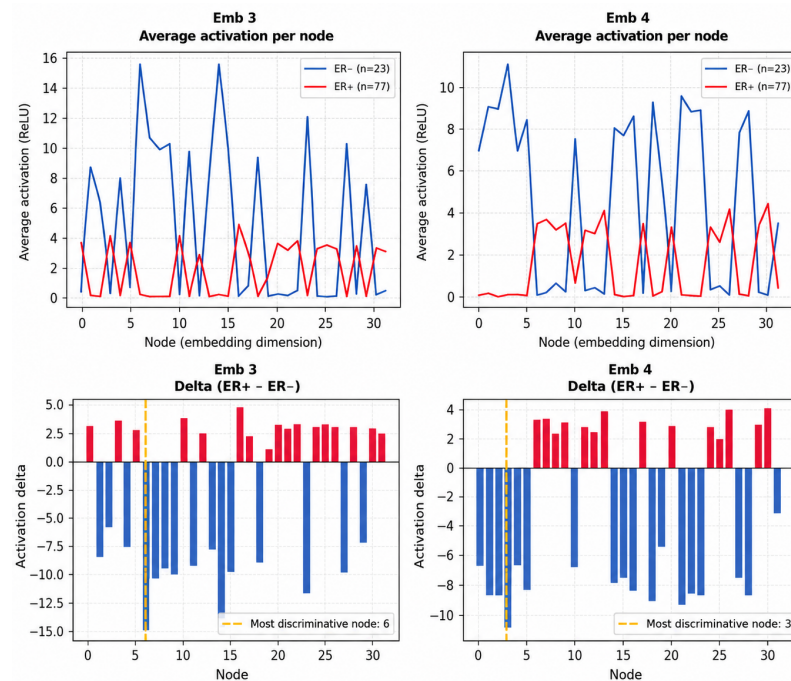


Figure 7. Average node activations for ER+ and ER− samples across the last parallel embedding branches of PLAT (top row), together with the activation differences between classes (bottom row). Positive delta values indicate stronger activation in ER+ samples, whereas negative values indicate higher activation in ER− samples. The highlighted nodes correspond to the most discriminative embedding dimensions.

The bottom row displays the activation delta, defined as ER+ minus ER−, for each latent node. Positive values correspond to nodes more strongly activated in ER+ samples, whereas negative values indicate higher activation in ER− patients. Across the four embedding branches (Figures 6 and 7), several latent nodes exhibit pronounced class-specific differences, demonstrating that different branches capture complementary discriminative information. These results further support the hypothesis that the proposed parallel embedding strategy distributes subtype-specific transcriptomic information across multiple latent representations rather than concentrating it within a single embedding branch.

The highlighted dimensions correspond to the most discriminative nodes within each embedding branch. These nodes represent latent features where the activation difference between ER+ and ER− samples is maximal, suggesting that they encode highly subtype-specific transcriptomic information. Interestingly, different embeddings emphasize different discriminative dimensions, supporting the hypothesis that the parallel embedding strategy encourages representation diversity and allows the model to capture complementary biological signals. Notably, these discriminative nodes consistently exhibit clearer separation for the ER− subtype, and their activation levels are generally higher in ER− samples despite this class being underrepresented in the dataset. This observation is consistent with the ability of the proposed architecture to learn informative latent representations for the minority class despite the moderate class imbalance present in the TCGA-BRCA dataset.

The analysis presented in Figure 8 investigates the genes associated with the most discriminative latent nodes identified in each parallel embedding branch of the proposed architecture (see Figures 6 and 7). For each embedding, the node with the highest activation delta, defined as the largest absolute difference in mean activations between classes, was selected. The genes with the largest absolute weights associated with that node were then

extracted. This procedure provides additional interpretability by identifying the biological signals most strongly related to the model's discriminative representations.

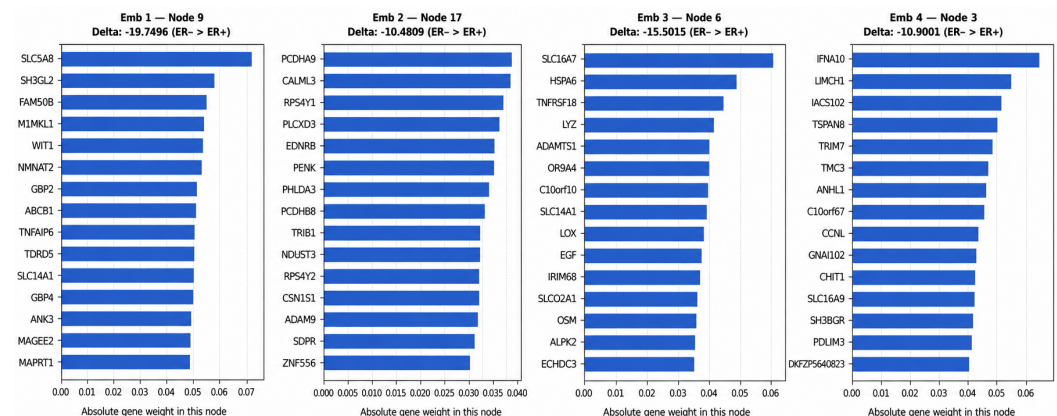


Figure 8. Top genes with the highest absolute weights associated with the most discriminative latent node of each embedding branch in the proposed architecture. Each panel highlights genes contributing most strongly to subtype separation between ER+ and ER− samples in the TCGA Breast Cancer dataset, suggesting that different embeddings capture complementary biological signals related to estrogen receptor status.

The results reveal that each embedding branch emphasizes a distinct subset of genes, suggesting that the parallel embedding strategy learns complementary transcriptomic representations rather than redundant feature spaces. Three embeddings (Emb1, Emb3, and Emb4) were enriched in genes related to interferon signaling, immune response, macrophage activity, extracellular matrix remodeling, and tumor aggressiveness (e.g., ABCB1 [21], GBP2, GBP4, IFNA10, LYZ, CHIT1, LOX, OSM [22] and TSPAN8 [23]), all processes frequently associated with ER− and basal-like breast cancer phenotypes. In contrast, Emb2 was characterized by genes involved in epithelial differentiation and cellular regulation [24], suggesting a more luminal-associated transcriptional program.

The biological interpretation of the genes enriched within the learned embeddings further supports the ability of the proposed architecture to capture meaningful molecular programs rather than isolated predictive features. Specifically, GBP2, GBP4, and IFNA10 are closely associated with interferon-mediated immune signaling, a hallmark of immune-activated breast tumors. Guanylate-binding proteins (GBPs) are interferon-inducible GTPases involved in innate immunity, antigen presentation, and anti-tumor immune responses. Previous studies have shown that elevated GBP expression is characteristic of immune-enriched basal-like and triple-negative breast cancer (TNBC) subtypes, where it forms part of interferon- γ gene signatures associated with enhanced immune infiltration and distinct clinical outcomes [25,26]. Likewise, type I interferon signaling, represented here by IFNA10, plays an important role in regulating the interaction between tumor cells and the immune microenvironment and has been implicated in aggressive breast cancer phenotypes [27].

A second functional module comprises LYZ, CHIT1, and OSM, genes that are predominantly associated with inflammatory and stromal components of the tumor microenvironment. Both LYZ and CHIT1 are well-established markers of activated myeloid cells and macrophages, reflecting innate immune activation and inflammatory infiltration within breast tumors. Oncostatin M (OSM), a cytokine primarily secreted by macrophages and other immune cells, promotes epithelial-to-mesenchymal transition (EMT), cancer stem cell plasticity, extracellular matrix remodeling, and metastatic progression through activation of the JAK/STAT signaling pathway. Elevated OSM signaling has been associated with

poor prognosis and increased tumor aggressiveness, particularly in ER-negative and triple-negative breast cancers, highlighting the contribution of inflammatory stromal interactions to disease progression.

The remaining genes are primarily involved in extracellular matrix organization, cell migration, and therapeutic resistance. LOX encodes lysyl oxidase, a key enzyme responsible for collagen cross-linking and extracellular matrix stiffening, thereby facilitating tumor invasion and metastatic dissemination. TSPAN8, a member of the tetraspanin family, regulates cell adhesion, migration, extracellular vesicle formation, and metastatic colonization, and its overexpression has been reported in both primary breast tumors and distant metastases [23]. Finally, ABCB1 is one of the best-characterized ATP-binding cassette transporters and is widely recognized as a major mediator of multidrug resistance through active efflux of chemotherapeutic agents, contributing to treatment failure in several breast cancer subtypes.

Interestingly, these genes do not converge on a single biological pathway but instead organize into complementary functional modules involving interferon-mediated immune activation, inflammatory remodeling of the tumor microenvironment, extracellular matrix organization, metastatic dissemination, and drug resistance. This functional diversity suggests that the parallel attention branches are learning complementary latent representations that encode distinct biological programs rather than redundant subsets of genes. Such behavior is consistent with the objective of the proposed architecture, which seeks to disentangle multiple molecular processes contributing to the ER-positive and ER-negative phenotypes while maintaining predictive performance and improving biological interpretability.

To further validate the biological interpretability of the learned embeddings, we performed an over-representation analysis (ORA) using three complementary annotation resources: Gene Ontology Biological Process (GO-BP), KEGG 2021 Human, and DisGeNET. Significant enrichment was consistently observed for pathways and biological processes involved in interferon signaling, cytokine-mediated immune responses, innate immunity, and NOD-like receptor signaling, while disease enrichment highlighted immune-related and inflammatory conditions. These results are in agreement with the functional roles of the most influential genes identified by the model, including GBP2, GBP4, IFNA10, LYZ, CHIT1, OSM, LOX, TSPAN8, and ABCB1, which collectively participate in immune activation, extracellular matrix remodeling, metastatic progression, and therapeutic resistance.

Importantly, these enriched processes are consistent with the well-established molecular characteristics of ER-negative/basal-like breast tumors, providing independent biological validation that the proposed parallel attention architecture learns complementary and biologically coherent latent representations rather than redundant subsets of predictive genes. This additional analysis therefore reinforces both the predictive capability and the biological interpretability of the proposed model. The complete enrichment results are summarized in Figure 9.

To further assess the interpretability of the proposed architecture, we additionally performed a SHAP-based feature importance analysis of the baseline MLP. The 300 genes with the highest mean absolute SHAP values were subsequently subjected to the same functional enrichment analyses as those applied to PLAT. Although PLAT and the baseline MLP identified largely different sets of predictive genes, as reflected by the low Jaccard index (0.0141) and Overlap coefficient (0.0847), the subsequent functional enrichment analyses led to similar biological conclusions. In both cases, the identified genes were significantly enriched for pathways and disease annotations associated with breast cancer. This agreement at the functional level, despite the limited overlap at the individual-gene

level, suggests that the two architectures capture complementary molecular signatures converging on common biological processes.

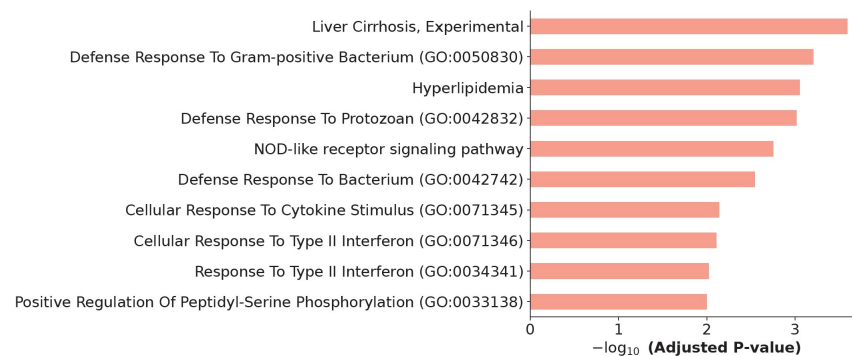


Figure 9. Over-representation analysis (ORA) of the genes identified by the proposed attention-based architecture. Significant functional enrichment (adjusted $p < 0.05$) was assessed using Gene Ontology Biological Process (GO-BP), KEGG 2021 Human, and DisGeNET.

4. Discussion

The present study explored the application of Transformer-inspired self-attention mechanisms to high-dimensional omic tabular data through a novel architecture based on parallel dense embeddings, referred to as PLAT (Parallel Latent Attention Transformer). The proposed approach was motivated by the hypothesis that explicitly modeling interactions between complementary latent representations could improve both predictive performance and interpretability in transcriptomic classification tasks.

The simulation experiments demonstrated that the behavior of the proposed architecture strongly depends on the signal-to-noise ratio and on the proportion of informative variables. In the Negative Binomial simulation framework, both the baseline MLP and the PLAT achieved their best performance at intermediate dimensionalities, whereas predictive accuracy decreased as the number of non-informative genes increased. This trend is consistent with the statistical challenges typically associated with omic datasets, where only a relatively small subset of genes contributes meaningful biological information while the majority of variables represent noisy background expression. Interestingly, the attention-based architecture showed its most favorable behavior in the 10% informative-feature scenario, suggesting the existence of an optimal representation regime in which self-attention mechanisms can more effectively identify informative feature interactions under moderate sparsity conditions.

An important contribution of this work lies in the analysis of lightweight attention-based architectures. The experiments revealed that substantial reductions in model complexity can be achieved by decreasing the number of parallel embeddings and latent dimensions while preserving competitive predictive performance. In particular, intermediate lightweight configurations maintained high AUC values despite using considerably fewer trainable parameters than the standard attention model. These findings are especially relevant in omic applications, where computational efficiency and parameter control become critical due to the large dimensionality and relatively limited sample sizes commonly encountered in transcriptomic studies.

Nevertheless, the results also indicated that excessive compression negatively affects stability and predictive capacity, particularly under NB-distributed data characterized by strong overdispersion and heterogeneous variance. This observation suggests that attention-based models require a minimum representational capacity to effectively capture complex transcriptomic patterns.

The evaluation on the TCGA-BRCA dataset further supported the effectiveness of the proposed architecture in a realistic biological setting. As summarized in Table 5, PLAT achieved predictive performance comparable to both the baseline MLP and the lightweight FT-Transformer across all evaluation metrics. Although the baseline MLP obtained slightly higher values for accuracy, balanced accuracy, sensitivity, and specificity, PLAT achieved the highest PR-AUC while exhibiting an AUC comparable to that of the MLP. In summary, the three architectures demonstrated similar predictive performance, with only modest differences across the evaluation metrics. The superior PR-AUC is particularly relevant given the moderate class imbalance of the TCGA-BRCA dataset (386 ER+ versus 113 ER− patients), as this metric better reflects the ability of the classifier to identify the minority class. These findings indicate that PLAT achieves a favorable trade-off between predictive performance and computational efficiency for high-dimensional transcriptomic classification, maintaining competitive accuracy with only a modest increase in model complexity. Equally important, the proposed architecture offers an interpretable representation of the learned features, as the attention analysis identifies biologically meaningful latent dimensions associated with subtype-specific transcriptomic signatures. The attention analysis revealed that the model distributes importance unevenly across the different embedding branches depending on the molecular subtype being analyzed. Certain embeddings received substantially higher attention in ER+ samples, whereas others specialized in ER− tumors, suggesting that the parallel embedding strategy encourages representation diversity and subtype-specific specialization within the latent space. This behavior supports the hypothesis that different embedding branches capture complementary transcriptomic patterns rather than redundant information.

The node activation analyses provided additional evidence regarding the internal organization of the learned latent representations. Several embedding dimensions exhibited pronounced class-specific activation patterns, suggesting that specific latent nodes play a prominent role in distinguishing between subtypes. Importantly, different embeddings emphasized different discriminative nodes, again supporting the idea that the architecture distributes biological information across multiple complementary latent subspaces. Such distributed representations may help improve robustness in high-dimensional biological problems where relevant signals are sparse and heterogeneous.

The gene-level interpretation analyses further reinforced the biological relevance of the learned representations. The most discriminative latent nodes were associated with distinct subsets of genes across embeddings, several of which are known to participate in cancer progression, immune response, or cellular signaling pathways. Interestingly, many of the strongest discriminative activations were associated with ER− tumors, potentially reflecting the greater heterogeneity and biological complexity of this subtype. These observations suggest that the proposed attention-based architecture is capable not only of improving predictive performance but also of identifying biologically meaningful transcriptomic structures associated with clinically relevant phenotypes.

Despite these promising results, several limitations should be acknowledged. First, the real-data evaluation was restricted to a single TCGA cohort and a binary ER classification problem. Additional validation across multiple omic datasets, cancer types, and multi-class classification tasks would be necessary to assess the generalizability of the proposed approach. Second, although the interpretability analyses provide insight into the internal behavior of the model, the biological significance of the identified latent representations and genes requires further experimental validation. Finally, while the proposed architecture remains relatively lightweight compared with larger Transformer models, scalability to extremely large transcriptomic cohorts and multi-omic integration scenarios should be investigated in future work.

In summary, the results suggest that self-attention over parallel dense embeddings constitutes a promising framework for high-dimensional omic data analysis. The architecture combines competitive predictive performance, parameter efficiency, and enhanced interpretability, offering an attractive alternative to conventional dense neural networks for transcriptomic classification problems.

5. Conclusions

This work introduced PLAT (Parallel Latent Attention Transformer), a self-attention architecture for high-dimensional omic tabular data based on parallel latent embeddings. The proposed architecture achieved competitive predictive performance in both Negative Binomial simulation studies and the TCGA-BRCA transcriptomic dataset when compared with a baseline multilayer perceptron and a lightweight FT-Transformer of comparable complexity. The simulation study further showed that the main advantage of PLAT is concentrated in specific high-dimensional settings with an intermediate proportion of informative features, while maintaining performance comparable to the alternative architectures across the remaining scenarios. Beyond predictive performance, the proposed architecture provides an interpretable representation of transcriptomic data through multiple parallel attention branches. Functional enrichment analyses demonstrated that these branches capture complementary biological programs associated with estrogen receptor status. Furthermore, a SHAP-based analysis of the baseline MLP showed that, although both models identify largely different sets of predictive genes, they converge to similar biological processes and disease pathways, supporting the biological relevance of the representations learned by PLAT. Overall, the results indicate that PLAT constitutes a competitive and interpretable framework for high-dimensional transcriptomic classification, combining parallel attention mechanisms with biologically meaningful latent representations while maintaining a model complexity comparable to existing deep learning architectures.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app16167890/s1>.

Author Contributions: Conceptualization, F.R.; methodology, K.E., N.J.G. and F.R.; software, K.E.; validation, K.E. and N.J.G.; formal analysis, F.R.; investigation, K.E. and N.J.G.; visualization, K.E.; writing—original draft preparation, F.R.; writing—review and editing, K.E., N.J.G., A.S.-P. and F.R.; supervision, A.S.-P., F.R. All authors have read and agreed to the published version of the manuscript.

Funding: Our research group has public funding (Ministerio de Ciencia, Innovación y Universidades, PID2023-148013OB-C22/AEI DOI:10.13039/501100011033) which finances all the group's production, including but not limited to this work.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The RNA-seq gene expression data analyzed in this study are publicly available from the Genomic Data Commons (GDC) Data Portal (<https://portal.gdc.cancer.gov/>) (accessed on 1 August 2026) under the TCGA-BRCA project. The source code and the generated simulated datasets required to reproduce the analyses presented in this study are available at <https://github.com/elatifikamal9-sudo/tabular-attention-rnaseq.git> (accessed on 1 August 2026).

Acknowledgments: ChatGPT Go 5.5 (OpenAI) was used as an assistive tool during manuscript preparation. Its use was limited to the generation of the architectural illustration shown in Figure 1 and to minor graphical formatting and layout adjustments of other figures. ChatGPT Go was not used to generate, modify, or interpret any experimental data, numerical values, statistical analyses, axis labels, data points, curves, or scientific conclusions. All computational analyses, visualizations, and results were produced by the authors and verified against the original outputs.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gorishniy, Y.; Kotelnikov, A.; Babenko, A. TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling. In *Proceedings of the International Conference on Learning Representations*; ICLR: Appleton, WI, USA, 2025.
2. O'Brien Quinn, H.; Sedky, M.; Francis, J.; Streeton, M. Literature Review of Explainable Tabular Data Analysis. *Electronics* **2024**, *13*, 3806. <https://doi.org/10.3390/electronics13193806>.
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; American Institute of Physics: College Park, MD, USA, 2017; Volume 30.
4. Niu, Z.; Zhong, G.; Yu, H. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing* **2021**, *452*, 48–62. <https://doi.org/10.1016/j.neucom.2021.03.091>.
5. Duru, I.; Sunar, A.S. Transformer and Pre-Transformer Model-Based Sentiment Prediction with Various Embeddings: A Case Study on Amazon Reviews. *Entropy* **2025**, *27*, 1202. <https://doi.org/10.3390/e27121202>.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*; ICLR: Appleton, WI, USA, 2025.
7. Zhao, S.; Wu, Y.; Tong, M.; Yao, Y.; Qian, W.; Qi, S. CoT-XNet: Contextual Transformer with Xception Network for Diabetic Retinopathy Grading. *Phys. Med. Biol.* **2022**, *67*, 245003. <https://doi.org/10.1088/1361-6560/ac9fa0>.
8. Arik, S.Ö.; Pfister, T. TabNet: Attentive Interpretable Tabular Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; PKP: Warsaw, Poland, 2021; Volume 35, pp. 6679–6687.
9. Somepalli, G.; Goldblum, M.; Schwarzschild, A.; Bruss, C.B.; Goldstein, T. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. In *Proceedings of the NeurIPS Workshop on Table Representation Learning*; American Institute of Physics: College Park, MD, USA, 2022.
10. Thielmann, A.; Reuter, A.; Säfken, B. Beyond Black-Box Predictions: Identifying Marginal Feature Effects in Tabular Transformer Networks. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research*; PMLR: New York, NY, USA, 2026; Volume 300.
11. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting Deep Learning Models for Tabular Data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943.
12. Huang, X.; Khetan, A.; Cvitkovic, M.; Karnin, Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. *arXiv* **2020**, arXiv:2012.06678.
13. Aggarwal, C.C. *Neural Networks and Deep Learning*; Springer: Cham, Switzerland, 2018.
14. Oberg, A.L.; Bot, B.M.; Grill, D.E.; Poland, G.A.; Therneau, T.M. Technical and Biological Variance Structure in mRNA-Seq Data: Life in the Real World. *BMC Genom.* **2012**, *13*, 304. <https://doi.org/10.1186/1471-2164-13-304>.
15. Cancer Genome Atlas Network. Comprehensive Molecular Portraits of Human Breast Tumours. *Nature* **2012**, *490*, 61–70. <https://doi.org/10.1038/nature11412>.
16. Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120. <https://doi.org/10.1038/ng.2764>.
17. Geyer, F.C.; Rodrigues, D.N.; Weigelt, B.; Reis-Filho, J.S. Molecular Classification of Estrogen Receptor-Positive/Luminal Breast Cancers. *Adv. Anat. Pathol.* **2012**, *19*, 39–53. <https://doi.org/10.1097/PAP.0b013e31823fafa0>.
18. Johnson, K.S.; Conant, E.F.; Soo, M.S. Molecular Subtypes of Breast Cancer: A Review for Breast Radiologists. *J. Breast Imaging* **2021**, *3*, 12–24. <https://doi.org/10.1093/jbi/wbaa110>.
19. Sheu, R.K.; Pardeshi, M.S. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors* **2022**, *22*, 8068. <https://doi.org/10.3390/s22208068>.
20. Zhang, K.; Wang, D.; Lin, F.; Xie, J.; Zhou, W. A Comprehensive Review of Explainable Artificial Intelligence in Healthcare: Methods, Evaluation, and Clinical Integration. *iScience* **2026**, *29*, 115026. <https://doi.org/10.1016/j.isci.2026.115026>.
21. Nedeljković, M.; Tanić, N.; Prvanović, M.; Milovanović, Z.; Tanić, N. Friend or Foe: ABCG2, ABCC1 and ABCB1 Expression in Triple-Negative Breast Cancer. *Breast Cancer* **2021**, *28*, 727–736. <https://doi.org/10.1007/s12282-020-01210-z>.
22. Segovia-Mendoza, M.; Morales-Montor, J. Immune Tumor Microenvironment in Breast Cancer and the Participation of Estrogen and Its Receptors in Cancer Physiopathology. *Front. Immunol.* **2019**, *10*, 348. <https://doi.org/10.3389/fimmu.2019.00348>.
23. Voglstaetter, M.; Thomsen, A.R.; Nouvel, J.; Koch, A.; Jank, P.; Navarro, E.G.; Gainey-Schleicher, T.; Khanduri, R.; Groß, A.; Rossner, F.; et al. Tspan8 Is Expressed in Breast Cancer and Regulates E-Cadherin/Catenin Signalling and Metastasis Accompanied by Increased Circulating Extracellular Vesicles. *J. Pathol.* **2019**, *248*, 421–437. <https://doi.org/10.1002/path.5281>.
24. Gandhi, N.; Oturkar, C.C.; Das, G.M. Estrogen Receptor-Alpha and p53 Status as Regulators of AMPK and mTOR in Luminal Breast Cancer. *Cancers* **2021**, *13*, 3612. <https://doi.org/10.3390/cancers13143612>.

25. Hunt, E.N.; Kopacz, J.P.; Vestal, D.J. Unraveling the Role of Guanylate-Binding Proteins (GBPs) in Breast Cancer: A Comprehensive Literature Review and New Data on Prognosis in Breast Cancer Subtypes. *Cancers* **2022**, *14*, 2794. <https://doi.org/10.3390/cancers14112794>.
26. Quintero, M.; Adamoski, D.; Reis, L.M.; Ascensão, C.F.R.; Oliveira, K.R.S.; Goncalves, K.A.; Dias, M.M.; Carazzolle, M.F.; Dias, S.M.G. Guanylate-Binding Protein-1 Is a Potential New Therapeutic Target for Triple-Negative Breast Cancer. *BMC Cancer* **2017**, *17*, 727. <https://doi.org/10.1186/s12885-017-3726-2>.
27. Provance, O.K.; Lewis-Wambi, J. Deciphering the Role of Interferon Alpha Signaling and Microenvironment Crosstalk in Inflammatory Breast Cancer. *Breast Cancer Res.* **2019**, *21*, 59. <https://doi.org/10.1186/s13058-019-1140-1>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.