Hyperparameter Tuning

We employed a systematic approach to determine the optimal hyperparameter configuration for the proposed model. The tuning process for the primary parameters was based on a grid search performed over a reasonable range of values, which was informed by both empirical data and established conventions in the field. The final hyperparameters were selected after a comprehensive analysis of the outcomes. The detailed parameters are as Table 1.1:

Table 1.1 Main Parameters of PepGraphormer

Category	Parameter Name	Parameter Value	Description
Build Graph	window_size	20	We explored values within the set {2, 3, 5, 10, 15, 20} and observed that the model's performance was relatively robust across this range. A larger window size of 20 was ultimately selected, as it allows for the capture of longer-range co-occurrence patterns without incurring a significant performance degradation on shorter peptides in our validation set.
ESM2 Feature Extraction	esm_model	esm2_t36_3B_UR50D	To leverage powerful sequence representations, we evaluated several models from the ESM-2 family, including the 35M, 150M, 650M, and 3B parameter variants. Our experiments revealed a clear trend of performance improvement with increasing model size. The esm2_t36_3B_UR50D, in particular, demonstrated superior performance. This is shown in Table 1.2.
	esm_repr_layer	36	We selected the final layer to extract representations, as it is generally considered to contain the most contextually rich and abstract features.
	esm_feat_dim	2560	This value is determined by the chosen esm_model_name (esm2_t36_3B_UR50D) and is not a tunable hyperparameter. It defines the dimensionality of the sequence and node feature space.
	esm_batch_size	16	16 was chosen as a conservative upper limit to ensure the process could run stably on GPUs with 24GB of VRAM without encountering out-of-memory errors, especially given the large embedding dimension of the 3B model.
GAT Model Configuration	gat_layers	2	We evaluated GAT architectures with {2, 3, 4} layers and found that a 2-layer configuration provided the optimal performance. Deeper models with 3 or 4 layers did not yield further improvements and exhibited early signs of over-smoothing, while also incurring a higher computational cost. This is shown in Table 1.3.
	gat_heads	4	We performed a grid search over {2, 4, 8, 16, 32} heads. While 8 heads also performed well, 4 heads provided a comparable MCC with fewer parameters, leading to a more efficient and faster-converging model. This is shown

			in Table 1.4.
	gat_hidden_dim	128	The hidden dimension of the GAT layer was selected from {64, 128, 256}. A dimension of 128 provided the best trade-off between model capacity and generalization.
Training Process Control	batch_size	64	The training batch size was selected from the set {16, 32, 64, 128, 256}. Our experiments indicated that the model's final performance was relatively insensitive to this parameter within this range. Consequently, a moderate value of 64 was chosen.
	learning rate	0.001	A relatively high learning rate of 0.001 was found to be effective. This value allowed for rapid initial convergence without causing instability in the training process. This is shown in Table 1.5.
	optimizer	Adam	The Adam optimizer was chosen as it is a standard, robust, and effective choice for training deep learning models.
	weight decay	2.5e-05	A moderate weight decay value of 0.0025 was applied to prevent overfitting by penalizing large model weights, which was found to be beneficial for generalization on the validation set. This is shown in Table 1.6.
	loss_func	CE	Cross-Entropy Loss is the standard and most appropriate loss function for binary classification tasks, which directly optimizes for improving the model's predictive accuracy.
	dropout	0.1	A relatively low dropout rate of 0.1 was found to be sufficient for regularization. This is shown in Table 1.7.
Model Fusion	gat_esm_m	0.8	This crucial fusion weight was determined by a grid search over [0, 1]. The significance of this finding is further elaborated in the subsequent experimental analysis.

Table 1.2 Performance Comparison of ESM-2 Models with Different Parameter Variants

ESM-2 Models	ACC	SE	SP	AUC	MCC
esm2_t12_35M_UR50D	0.709	0.598	0.820	0.749	0.429
esm2_t30_150M_UR50D	0.720	0.669	0.771	0.761	0.443
esm2_t33_650M_UR50D	0.749	0.592	0.906	0.816	0.524
esm2_t36_3B_UR50D	0.754	0.592	0.915	0.818	0.536

Table 1.3 Analyzing the Impact of Gat_layers

Gat_layers	ACC	SE	SP	AUC	MCC	Times(s)
2	0.752	0.612	0.891	0.800	0.524	2753.57
3	0.743	0.563	0.923	0.800	0.520	3988.00
4	0.520	0.077	0.962	0.773	0.084	7649.29

Table 1.4 Analyzing the Impact of Gat_heads

Gat_heads	ACC	SE	SP	AUC	MCC	Times(s)
2	0.746	0.663	0.830	0.796	0.500	2617.72
4	0.752	0.612	0.891	0.800	0.524	2753.57
8	0.740	0.555	0.924	0.795	0.516	3636.29
16	0.743	0.552	0.933	0.804	0.525	5531.41
32	0.500	1.000	0.000	0.512	0.000	8605.42

Table 1.5 Analyzing the Impact of Learning Rate

Learning Rate	ACC	SE	SP	AUC	MCC
0.00001	0.872	0.771	0.974	0.963	0.761
0.0005	0.910	0.893	0.927	0.967	0.820
0.001	0.911	0.895	0.927	0.968	0.822
0.005	0.900	0.885	0.915	0.958	0.800

Table 1.6 Analyzing the Impact of Weight Decay

Weight Decay	ACC	SE	SP	AUC	MCC
2.5e-02	0.857	0.893	0.822	0.929	0.716
2.5e-03	0.900	0.885	0.915	0.958	0.800
2.5e-04	0.920	0.891	0.949	0.971	0.842
2.5e-05	0.925	0.903	0.947	0.972	0.851
2.5e-06	0.916	0.870	0.962	0.971	0.835
0	0.920	0.897	0.943	0.970	0.841

Table 1.7 Analyzing the Impact of Dropout

Dropout	ACC	SE	SP	AUC	MCC
0.05	0.920	0.891	0.949	0.971	0.842
0.1	0.922	0.895	0.949	0.973	0.845
0.2	0.910	0.927	0.893	0.967	0.820
0.3	0.904	0.921	0.887	0.965	0.808
0.5	0.897	0.854	0.939	0.963	0.796

Based on the above experiments, the optimal hyperparameters were selected as follows: ESM-2 model (esm2_t36_3B_UR50D), GAT layers (2), attention heads (4), learning rate (0.001), weight decay (2.5e-05), and dropout (0.1). The best performance for each parameter is highlighted in bold in the corresponding tables. These empirically validated choices collectively contribute to the superior performance of PepGraphormer.