Multimodal Alignment Reveals Interpretable Gene–Morphology Links in Perineuronal Net Pathology

Anonymous Author(s)
Affiliation
Address
email

Abstract

Perineuronal nets are extracellular matrix structures that enmesh specific neurons, and their disruption has been linked to glioma progression and epilepsy. Yet most studies analyze pathology images, gene expression, or clinical variables in isolation, limiting our understanding of how perineuronal net changes connect to disease. We present a joint multimodal framework that learns aligned embeddings from three inputs: pathology images, RNA expression, and clinical covariates, using a contrastive objective with cross-modality reconstruction and pathway-informed regularization. The approach supports missing modalities via modality dropout and gated fusion at inference, and provides interpretability through pathway enrichment analyses and attention maps that highlight morphology consistent with perineuronal net biology. On a small, patient-level multimodal cohort, the method outperforms early/intermediate/late fusion and unimodal baselines and yields transparent gene–morphology associations, suggesting a practical route to integrating limited multimodal data for perineuronal net pathology.

1 Introduction

Perineuronal nets (PNNs)[8, 14, 5] are specialized extracellular matrix (ECM) structures that enwrap neuronal somata and proximal dendrites, supporting synaptic stabilization and constraining plasticity. Composed of a hyaluronan backbone, chondroitin sulfate proteoglycans (e.g., aggrecan, brevican), link proteins (HAPLN family), and tenascin-R, PNNs stabilize perisomatic synapses. In glioma, PNN degradation has been associated with tumor invasion and seizure susceptibility[13], suggesting potential clinical utility.

However, prevailing assessments rely on subjective histopathology and seldom connect morphology to underlying molecular programs. Recent unimodal advances[25, 1] can reduce manual effort but are not designed to integrate complementary signals that jointly characterize PNN remodeling. By nature, PNN pathology is multimodal: immunofluorescence (IF) images capture mesoscale ECM and perisomatic changes[29]; transcriptomics reflects pathway shifts in matrix assembly, proteoglycan turnover, and synaptic stabilization; and clinical variables encode disease stage, treatment exposure, and seizure phenotype[26, 17, 23]. Considering any single view in isolation obscures gene–morphology associations and hampers clinical interpretability.

We propose a PNN-centric multimodal learning framework that jointly learns from IF images, RNA expression, and clinical covariates. The model aligns patient-level embeddings across modalities via a contrastive objective with pathway-informed regularization, while preserving modality-specific signals through unimodal reconstruction. To make gene–morphology associations explicit, we further include cross-modality reconstruction (e.g., image→RNA), which supports interpretability and hypothesis generation.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

We summarize our contributions as follows:

- We develop a PNN-centric multimodal framework that aligns patient-level *IF images*, transcriptomic, and clinical data via contrastive learning, jointly optimized with unimodal and cross-modality reconstruction objectives.
- We leverage cross-modality reconstruction together with attention/saliency readouts to analyze *gene–morphology associations* in a unified embedding space, while preserving modality-specific signals.
- Evaluated on a class-imbalanced glioma cohort under a full-modality setting, the framework learns a shared representation that co-clusters patients with concordant molecular–morphological–clinical profiles and improves over early-, intermediate-, and late-fusion as well as attention-based baselines on seizure prediction.

2 Related Work

Multimodal Learning in Healthcare and Pathology Large-scale vision–language models trained on paired medical images and reports have improved diagnostic performance and data efficiency [16, 18, 4]. Beyond text–image pairing, multimodal fusion of images with structured clinical variables improves risk prediction in diverse settings, such as pulmonary embolism [6], while specialized architectures integrate heterogeneous EHR signals (notes, labs, vitals) [24, 21]. In computational pathology, curated knowledge can guide pretraining and adaptation [31]. Contrastive objectives are effective for medical representation learning from paired or unpaired data [30, 28].

Histology–Genomics Integration A line of work studies joint modeling of histology and genomics for cancer diagnosis and prognosis, including fusion frameworks and co-attention designs that couple slide-level features with molecular readouts [10, 9, 11]. These methods motivate learning shared representations that preserve modality-specific signals while enabling cross-modal alignment—an idea we adapt to immunofluorescence (IF) images (tissue-level fluorescence microscopy of PNN markers), transcriptomics, and clinical variables in the context of perineuronal net pathology.

PNNs and Glioma-Related Epileptogenesis PNNs are specialized ECM structures implicated in synaptic stabilization and plasticity regulation. Imaging and mapping studies have quantified PNN organization (e.g., WFA-positive meshes and PV colocalization) [25, 23], while molecular analyses linked gene signatures to seizure phenotypes in glioma [19, 20]. MRI-based prediction of glioma-associated epilepsy has also been explored [27]. Our work complements these by aligning IF images with pathway-informed transcriptomic signals and clinical covariates to surface interpretable gene–morphology associations in PNN pathology.

3 Proposed Method: Multimodal Joint Training

In this section, we formalize the PNN-detection task and present a multimodal joint training framework that enhances predictive interpretability. An overview of the architecture is shown in Fig. 1.

3.1 Experiment Setup

We aim to address multimodal representation learning for clinical outcome prediction in PNN detection in this work. Formally, we denote the dataset as $\mathcal{D} = \{(I_i, R_i, C_i, Y_i)\}_{i=1}^{N}$ where each sample $d_i \in \mathcal{D}$ corresponds to a single *patient*; ROI-level features are aggregated per patient to prevent leakage across folds. Specifically, $I_i \in \mathbb{R}^{H\times W\times 3}$ represents an *immunofluorescence (IF)* image (or tile aggregate), $R_i \in \mathbb{R}^G$ denotes bulk/spatial RNA sequencing data with G genes, C_i indicates structured clinical covariates, and $Y_i \in \{0, 1\}$ denotes the binary seizure label. Our objective is to learn a unified embedding space that effectively integrates these heterogeneous modalities, thereby enabling a comprehensive cross-modal understanding and improving predictive performance for clinical outcomes.

3.2 Joint Training Framework

Given the heterogeneity of input modalities, we employ self-supervised objectives to train modality-specific encoders, combining reconstruction losses with cross-modal reconstruction and a contrastive objective applied to the bottleneck representations. Throughout, we refer to the d-dimensional encoder outputs $\mathbf{z}_I = f_I(I)$, $\mathbf{z}_R = f_R(R)$, and $\mathbf{z}_C = f_C(C)$ as the *bottleneck embeddings* (shared latent representations in $\mathbb{R}^d$) used by downstream reconstruction, cross-modality projection, and contrastive

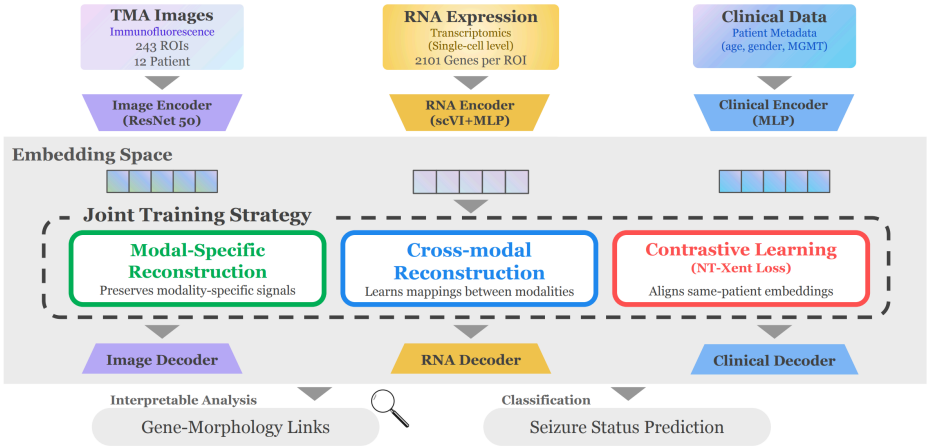

Figure 1: Overview of the multimodal joint training framework for PNN-centric glioma analysis. This framework encodes immunofluorescence (IF) images, RNA profiles (bulk/spatial; not single-cell), and clinical data via modality-specific encoders, jointly trained with *unimodal reconstruction* (Green), *cross-modality reconstruction* (Blue), and a *contrastive* objective (Red). The resulting embeddings support seizure prediction and interpretable gene–morphology analyses.

objectives. This design ensures that each embedding retains *modality-dependent characteristics* while remaining *well-aligned* within a shared latent space to support cross-modal integration.

3.2.1 Reconstruction Learning Strategy

To capture information across different modalities, each input is first compressed into a latent representation by its corresponding encoder and then reconstructed back into its original space through a dedicated decoder. We adopt distinct encoding strategies tailored to each data modality:

- **Image Encoder.** We use a ResNet-50[15] architecture (with the final classification layer removed) to encode histopathology images. This encoder extracts high-level visual features and projects each image into a latent representation vector.
- **RNA Encoder.** We use a VAE-based encoder (scVI-style latent model[22]) adapted to bulk/spatial RNA profiles for dimensionality reduction. An MLP g_{θ_R} maps the latent to the shared embedding:
$$\mathbf{z}_R = g_{\theta_R}(\text{scVI}(R)) \in \mathbb{R}^d.$$
- **Clinical Encoder.** Categorical clinical variables are encoded using learnable embedding layers. For the k-th clinical variable, an encoder $\mathbf{Enc}_k$ produces its representation, which is then aggregated with importance weighting:
$$\mathbf{z}_C = \sum_k w_k \cdot \mathbf{Enc}_k(C_k),$$
where w_k denotes the *learnable* importance weight associated with the k-th clinical variable.

Each modality-specific decoder ($\mathbf{Dec}_I, \mathbf{Dec}_R, \mathbf{Dec}_{C_k}$) subsequently reconstructs the original input from its latent embedding, producing outputs $\hat{I}$, $\hat{R}$, and $\hat{C}$. Clinical reconstruction uses a shared clinical decoder with per-variable lightweight heads $\mathbf{Dec}_{C_k}$ that are shared across all patients; we keep the per-variable notation for clarity. The joint reconstruction objective is formulated as: $\mathcal{L}_{\text{recon}} = \text{MSE}(\hat{I}, I) + \text{MSE}(\hat{R}, R) + \sum_k \text{CrossEntropy}(\hat{C}_k, C_k)$. This strategy yields informative, modality-specific embeddings. Given the inherent correlations across modalities, we further introduce **Cross-Modality Reconstruction** objectives to induce semantically meaningful cross-modal mappings. Specifically, we employ modality-to-modality projection networks to ensure dimensional consistency and to align representations within a shared high-dimensional space. For example, we use MLPs to project representations into the space of a target modality:

$$\mathbf{z}_{I\to R} = \text{MLP}_{\theta_{I\to R}}(\mathbf{z}_I), \quad \mathbf{z}_{R\to I} = \text{MLP}_{\theta_{R\to I}}(\mathbf{z}_R)$$

where $\text{MLP}_{\theta_{I\to R}}$ and $\text{MLP}_{\theta_{R\to I}}$ denote the projection networks from image to RNA space and from RNA to image space, respectively. Analogous to $\mathcal{L}_{\text{recon}}$, the cross-modal reconstruction loss is defined

as:

$$\mathcal{L}_{\text{cross}} = \text{MSE}(\mathbf{Dec}_R(\mathbf{z}_{I\to R}), R) + \text{MSE}(\mathbf{Dec}_I(\mathbf{z}_{R\to I}), I) + \sum_k \text{CrossEntropy}(\mathbf{Dec}_{C_k}(\mathbf{z}_{I\to C}), C_k)\,.$$

This joint paradigm promotes coherent representations across heterogeneous modalities, facilitating integration in a shared embedding space.

3.2.2 Contrastive Learning Strategy

To promote alignment of representations across modalities, we also incorporate a contrastive learning objective that enforces embeddings derived from the same patient to keep close in representation space while ensuring separability across patients. Specifically, for each patient i, the modality-specific embeddings $(\mathbf{z}_I^i, \mathbf{z}_R^i, \mathbf{z}_C^i)$ are treated as positive pairs, whereas embeddings from different patients $j \neq i$ serve as negatives. We apply the NT-Xent loss [12] to all modality pairs: $\mathcal{L}_{\text{contrast}} = \mathcal{L}_{I,R} + \mathcal{L}_{R,C} + \mathcal{L}_{C,I}$, where each pairwise contrastive loss is defined as

$$\mathcal{L}_{m_1,m_2} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(\mathbf{z}_{m_1}^i, \mathbf{z}_{m_2}^i)/\tau\right)}{\sum_{j=1}^{N} \exp\Big(\text{sim}(\mathbf{z}_{m_1}^i, \mathbf{z}_{m_2}^j)/\tau\Big)}$$

with $\text{sim}(\cdot,\cdot)$ denoting cosine similarity, τ the temperature parameter, and N the batch size. Building on the proposed contrastive and reconstruction strategies, the final training objective jointly optimizes all three components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \lambda_{\text{contrast}}\,\mathcal{L}_{\text{contrast}} + \lambda_{\text{cross}}\,\mathcal{L}_{\text{cross}},$$

where λ_{recon}, $\lambda_{\text{contrast}}$, and λ_{cross} balance the losses. We select $(\lambda_{\text{recon}}, \lambda_{\text{contrast}}, \lambda_{\text{cross}})$ by grid search on training folds with early stopping and fix the chosen triplet across folds; the final values are $(0.5,\ 1.0,\ 0.5)$. This joint objective enables the model to capture biologically grounded cross-modal correspondences, yielding a representation space in which patients with similar molecular profiles, tissue morphology, and clinical features cluster coherently. Such alignment facilitates transparent and clinically relevant interpretation across modalities, even under data-limited conditions.

4 Experiment Result

4.1 Dataset & Metric

Dataset. Our cohort comprises 12 glioma patients (4 low-grade glioma, LGG; 8 glioblastoma, GBM), with 243 tissue microarray (TMA) regions of interest (ROIs). Each ROI includes paired immunofluorescence images, RNA expression profiles (2,101 genes), and clinical metadata. Further dataset details are provided in Appendix A.1.

Metrics. We formulate the prediction of the seizure status as a binary classification problem, where each sample i is assigned a predicted score $\hat{y}_i \in [0, 1]$. To prevent data leakage, we perform **patient-wise** K-fold cross-validation, ensuring that all ROIs from the same patient are assigned to the same fold. We report the average **Accuracy**, **Precision**, **Recall**, **F1**, and **AUC** (area under the receiver operating characteristic curve) across folds.

4.2 Evaluation

We compare our approach against five standard modality-fusion paradigms:, Early Fusion[10, 11], Late Fusion[7], Intermediate Fusion[10], Attention-based Fusion[9], and Cross-modal Transformer[9]. Detailed descriptions of each method are provided in Appendix A.2.For fair comparison, we use the same modality-specific encoders and apply the background removal method to minimize the impact of background noise. The results are presented in Table 1

Naive fusion baselines (feature concatenation or weighted averaging) underperform because they ignore cross-modal heterogeneity and differences in scale and structure. Attention-based fusion partially mitigates this by modeling interactions, yet it still lacks explicit patient-level alignment and does not preserve modality-specific signal, yielding limited gains in data-limited settings. In contrast, our pathway-aware contrastive training with cross-modal reconstruction simultaneously enforces cross-modal alignment and preserves modality-specific representations. This leads to consistently superior performance across all evaluation metrics, thereby validating the effectiveness of our proposed approach.

Table 1: Performance comparison of fusion methods under identical encoders, splits, and preprocessing. Our joint contrastive training attains the best overall performance.

	Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**
Early Fusion(Concat)	0.713	0.736	0.698	0.716	0.784
Late Fusion(Weighted Average)	0.738	0.754	0.721	0.737	0.802
Intermediate Fusion	0.759	0.772	0.745	0.762	0.821
Attention-based Fusion	0.782	0.798	0.769	0.785	0.845
Cross-Modal Transformer	0.796	0.811	0.783	0.798	0.856
Joint Training (Ours)	**0.824**	**0.841**	**0.812**	**0.826**	**0.887**

4.3 Interpretability Analysis

To assess the biological mechanisms driving our model's predictions and establish the clinical relevance of the learned representations, we conduct comprehensive interpretability analyses, including cross-modal attention visualization, region-specificity assessment, and gene–morphology association mapping. These analyses indicate that the model captures biologically meaningful patterns consistent with established PNN biology and glioma pathophysiology. Full procedures and results are provided in Appendix A.4.

5 Conclusion

We presented a multimodal framework for PNN-centric glioma analysis that couples biologically informed contrastive alignment with cross-modal and unimodal reconstruction, enabling joint prediction over histopathology, RNA expression, and clinical metadata. On a 12-patient (GBM/LGG) cohort with 243 ROIs and patient-level cross-validation, the proposed model consistently outperforms strong fusion baselines, indicating that shared representation learning captures cross-modal dependencies more effectively than post-hoc score fusion or attention alone. The learned embeddings also support clinically meaningful interpretation by linking histomorphology to transcriptomic signatures.

References

[1] Jordan T Ash, Gregory Darnell, Daniel Munro, and Barbara E Engelhardt. Joint analysis of expression levels and histological images identifies genes associated with tissue morphology. *Nature communications*, 12(1):1609, 2021.

[2] <Update authors here>. Healnet: Learning with incomplete modalities in healthcare. <*Update venue here*>, <YYYY>. URL `<optional>`. Please update with the correct citation (DOI/arXiv/venue).

[3] <Update authors here>. Multimodn: A missing-modality robust framework for clinical multimodal learning. <*Update venue here*>, <YYYY>. URL `<optional>`. Please update with the correct citation (DOI/arXiv/venue).

[4] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

[5] Ivan Banovac, Matija Vid Prkačin, Ivona Kirchbaum, Sara Trnski-Levak, Mihaela Bobić-Rasonja, Goran Sedmak, Zdravko Petanjek, and Natasa Jovanov-Milosevic. Morphological and molecular characteristics of perineuronal nets in the human prefrontal cortex—a possible link to microcircuitry specialization. *Molecular neurobiology*, 62(1):1094–1111, 2025.

[6] Noa Cahan, Eyal Klang, Edith M Marom, Shelly Soffer, Yiftach Barash, Evyatar Burshtein, Eli Konen, and Hayit Greenspan. Multimodal fusion models for pulmonary embolism mortality prediction. *Scientific Reports*, 13(1):7544, 2023.

[7] Francisco Carrillo-Perez, Juan Carlos Morales, Daniel Castillo-Secilla, Olivier Gevaert, Ignacio Rojas, and Luis Javier Herrera. Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. *Journal of Personalized Medicine*, 12(4): 601, 2022. doi: 10.3390/jpm12040601.

[8] Marco R Celio, Roberto Spreafico, Silvia De Biasi, and Laura Vitellaro-Zuccarello. Perineuronal nets: past and present. *Trends in neurosciences*, 21(12):510–515, 1998.

[9] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew F K Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021.

[10] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew F K Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022. doi: 10.1109/TMI.2020.3021387.

[11] Richard J Chen, Ming Y Lu, Drew F K Williamson, Tiffany Y Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878.e6, 2022. doi: 10.1016/j.ccell.2022.07.004.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[13] Egor Dzyubenko, Katrin I Willig, Dongpei Yin, Maryam Sardari, Erdin Tokmak, Patrick Labus, Ben Schmermund, and Dirk M Hermann. Structural changes in perineuronal nets and their perforating gabaergic synapses precede motor coordination recovery post stroke. *Journal of Biomedical Science*, 30(1):76, 2023.

[14] James W Fawcett, Toshitaka Oohashi, and Tommaso Pizzorusso. The roles of perineuronal nets and the perinodal extracellular matrix in neuronal function. *Nature Reviews Neuroscience*, 20 (8):451–465, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023.

[17] Sakina P Lemieux, Varda Lev-Ram, Roger Y Tsien, and Mark H Ellisman. Perineuronal nets and the neuronal extracellular matrix can be imaged by genetically encoded labeling of hapln1 in vitro and in vivo. *Biorxiv*, 2023.

[18] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[19] Jinwei Li, Shengrong Long, Yang Zhang, Wei Wei, Shuangqi Yu, Quan Liu, Xuhui Hui, Xiang Li, and Yinyan Wang. Molecular mechanisms and diagnostic model of glioma-related epilepsy. *NPJ Precision Oncology*, 8(1):223, 2024.

[20] Zesheng Li, Ting Tang, Ziqian Yan, Yongchang Lu, Mingshan Liu, Hongyi Huang, Penghu Wei, and Guoguang Zhao. Leveraging pathological markers of lower grade glioma to predict the occurrence of secondary epilepsy, a retrospective study. *Scientific Reports*, 15(1):23907, 2025.

[21] Zichen Liu, Xuyuan Liu, Yanlong Wen, Guoqing Zhao, Fen Xia, and Xiaojie Yuan. Treeman: Tree-enhanced multimodal attention network for icd coding. *arXiv preprint arXiv:2305.18576*, 2023.

[22] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[23] Leonardo Lupori, Valentino Totaro, Sara Cornuti, Luca Ciampi, Fabio Carrara, Edda Grilli, Aurelia Viglione, Francesca Tozzi, Elena Putignano, Raffaele Mazziotti, et al. A comprehensive atlas of perineuronal net distribution and colocalization with parvalbumin in the adult mouse brain. *Cell Reports*, 42(7), 2023.

[24] Weimin Lyu, Xinyu Dong, Rachel Wong, Songzhu Zheng, Kayley Abell-Hart, Fusheng Wang, and Chao Chen. A multimodal transformer: Fusing clinical notes with structured ehr data for

interpretable in-hospital mortality prediction. In *AMIA Annual Symposium Proceedings*, volume 2022, page 719, 2023.

[25] Mikhail Paveliev, Anton A Egorchev, Foat Musin, Nikita Lipachev, Anastasiia Melnikova, Rustem M Gimadutdinov, Aidar R Kashipov, Dmitry Molotkov, Dmitry E Chickrin, and Albert V Aganov. Perineuronal net microscopy: From brain pathology to artificial intelligence. *International Journal of Molecular Sciences*, 25(8):4227, 2024.

[26] Arnaud Tanti, Claudia Belliveau, Corina Nagy, Malosree Maitra, Fanny Denux, Kelly Perlman, Frank Chen, Refilwe Mpai, Candice Canonne, Stéphanie Théberge, et al. Child abuse associates with increased recruitment of perineuronal nets in the ventromedial prefrontal cortex: a possible implication of oligodendrocyte progenitor cells. *Molecular psychiatry*, 27(3):1552–1561, 2022.

[27] Wei Wang, Xuanyi Li, Lou Ye, and Jian Yin. A novel deep learning model for glioma epilepsy associated with the identification of human cytomegalovirus infection injuries based on head mr. *Frontiers in Microbiology*, 14:1291692, 2023.

[28] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[29] AnnaLin M Woo, Erik J Fleischel, Dipan C Patel, and Harald Sontheimer. Contribution of perineuronal nets to hyperexcitability in pilocarpine-induced status epilepticus. *Epilepsia*, 2025.

[30] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 5th Conference on Medical Imaging with Deep Learning (MIDL)*, volume 182 of *Proceedings of Machine Learning Research*. PMLR, 2022.

[31] Xiao Zhou, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pretraining for computational pathology. In *European Conference on Computer Vision*, pages 345–362. Springer, 2024.

A Appendix

A.1 Dataset Detail

We present patient characteristics in Table 2. All analyses in the main paper use a 12-patient glioma subset; counts elsewhere refer to regions of interest (ROIs). The cohort comprises 12 patients with a mean age of 65.5 ± 13.2 years and a balanced gender distribution. Seizures were observed in 58.3% of patients, and the majority (66.7%) were diagnosed with glioblastoma.

Table 2: Patient demographics and clinical characteristics. The cohort comprises 12 patients with mixed diagnoses, including 8 glioblastomas and 4 lower-grade gliomas.
Diagnosis: IDs GBM1–8 are GBM; IDs starting with O/A are LGG.

Patient ID	Age	Gender	Brain Region	Seizures	ASM	Survival (months)
GBM1	75	F	Temporal	Yes	Keppra	20.5
GBM2	81	F	Frontal	No	None	3.5
GBM3	64	M	Temporal	Yes	Keppra	43.0
GBM4	78	M	Temporal	No	Keppra	10.0
GBM5	67	M	Frontal	No	None	16.0
GBM6	73	M	Frontoparietal	Yes	None	30.0
GBM7	84	F	Frontoparietal	Yes	None	1.0
GBM8	62	M	Temporal	Yes	None	8.0
O3_1	42	F	Frontal	Yes	Keppra	–
O2	47	F	Frontal	Yes	Keppra	–
A2	58	F	Frontal	No	None	–
O3_2	55	F	Frontal	No	None	–

A.1.1 Comprehensive Tissue Microarray Design and Quality Control

Our study utilized two tissue microarrays employing NanoString GeoMx Digital Spatial Profiling technology: TMA_1054 (whole transcriptome analysis, 190 initial ROIs) and TMA_1054B (whole transcriptome + protein panel, 285 initial ROIs). Each ROI measured 250 μm in diameter and underwent whole transcriptome analysis (WTA; 18,677 RNAs), with TMA_1054B additionally incorporating an immune profiling assay (IPA; $\sim$570 proteins). We implemented a six-step quality-control pipeline that reduced the initial 475 ROIs to 449 high-quality regions (176 from TMA_1054, 273 from TMA_1054B), representing $\sim$20% data reduction while preserving analytical integrity: (1) morphology/staining QC; (2) RNA integrity and cellularity checks; (3) segment filtering (10% threshold); (4) target filtering (10% threshold); (5) Q3 normalization; and (6) statistical-analysis preparation. This process refined the transcriptomic dataset from 18,677 to 16,561 reliable targets.

A.1.2 Spatial Compartment Classification and Distribution

From the quality-controlled dataset, we focused on 243 ROIs from 12 glioma patients (GBM and LGG) with complete clinical annotations. ROIs were classified into three spatial compartments based on tumor infiltration: Brain (B, n=83, 34.2%; 0% tumor content), Intermediate (I, n=80, 32.9%; $< 50\%$ infiltration), and Tumor (T, n=80, 32.9%; $> 50\%$ malignant cell content). The anatomical distribution encompassed 156 frontal (64.2%), 62 temporal (25.5%), and 25 frontoparietal (10.3%) regions. Seizure analysis identified 108 seizure-positive regions (44.4%) versus 135 seizure-negative (55.6%).

A.1.3 Molecular Profiling and PNN Quantification

Each ROI underwent dual-modal analysis combining immunofluorescence imaging and spatial transcriptomics. Immunofluorescence panels included perineuronal net markers (aggrecan staining), nuclear markers (SYTO13), and cellular markers (PanCK for epithelial cells). PNN abundance was quantified across spatial compartments, revealing progressive loss from Brain (highest) $\rightarrow$ Intermediate $\rightarrow$ Tumor (lowest) regions; the PNN loss metric (Brain% $-$ Tumor%) correlated with seizure frequency across tumor subtypes. The transcriptomic component captured spatial gene-expression patterns across the invasion spectrum, enabling identification of differentially expressed

genes between compartments and seizure states. This multimodal dataset supports analysis of gene–morphology relationships underlying PNN remodeling and glioma-associated epileptogenesis.

A.2 Baseline Methods

We compare against five standard fusion paradigms, instantiated on top of the same modality-specific encoders to ensure fairness:

- **Early Fusion (Concatenation)[10, 11].** L2-normalized embeddings from each modality are concatenated and fed to an MLP classifier; this setting is widely adopted in histopathology + omics pipelines.
- **Late Fusion (Weighted Average)[7].** Independent per-modality predictors are trained; calibrated probabilities (or logits) are combined via a validation-tuned weighted average.
- **Intermediate Fusion[10].** To capture cross-modal interactions beyond concatenation, we fuse features at a hidden layer with a tensor/bilinear-style head (Pathomic-Fusion–style).
- **Attention-based Fusion[9].** A gating/co-attention module learns modality importances and mixes features adaptively before classification (e.g., co-attention over WSI features and omics embeddings).
- **Cross-modal Transformer[9].** Directional cross-modal attention aligns and exchanges information across modalities prior to prediction (MCAT-style co-attention).

A.3 Ablation Study on Multi-Modality Training

To demonstrate the necessity of incorporating all modalities for PNN detection, we compare unimodal baselines against the proposed multimodal framework, as shown in Table 3.

Table 3: Performance comparison of unimodal and multimodal configurations. The proposed multimodal model achieves the best overall performance across all metrics.

	Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC**
Image Only (ResNet50)	0.752	0.778	0.734	0.755	0.821
RNA Only (scVI)	0.698	0.721	0.689	0.705	0.765
Clinical Only (XGBoost)	0.645	0.662	0.631	0.646	0.712
Multi-modal (Proposed)	**0.824**	**0.841**	**0.812**	**0.826**	**0.887**

The results show that the joint learning approach consistently outperforms single-modality models across all evaluation metrics. This highlights the critical role of integrating histopathology images, RNA profiles, and clinical data, where the complementary information from each modality enables more accurate and robust detection than relying on any individual source alone.

A.4 Interpretability Analysis

Understanding the biological mechanisms underlying model's predictions is crucial for clinical translation and scientific validation. We conducted comprehensive interpretability analyses to demonstrate how our multimodal framework learns clinically meaningful cross-modal relationships and captures known neurobiological patterns relevant to glioma-associated epilepsy.

A.4.1 Cross-Modal Attention Mechanisms

Our joint training framework learns interpretable attention patterns that reveal biologically meaningful associations between histological features and molecular pathways. As summarized in Fig. 2, *PNN_density* aligns most strongly with GABAergic signaling (High), *Tumor_boundary* with Proliferation (High), *Vessel_proximity* with Inflammatory signaling (Medium), and *Cell_density* with Metabolic pathways (Low). These patterns are consistent with PNN-mediated inhibitory regulation and invasion-associated proliferative signaling in glioma. *Importantly, Fig. 2 shows attention weights—not regional expression levels or brain-region comparisons.*

Histological Features	GABAergic Signaling	Inflammatory Response	Metabolic Activity	Proliferation Markers	Vascular Development	Synaptic Plasticity
PNN Density	0.92	0.34	0.45	0.23	0.31	0.78
Vessel Proximity	0.28	0.72	0.38	0.41	0.85	0.33
Cell Density	0.45	0.56	0.67	0.52	0.29	0.41
Tumor Boundary	0.19	0.83	0.35	0.88	0.64	0.22
GFAP Intensity	0.38	0.91	0.42	0.37	0.28	0.45
Nuclear Morphology	0.25	0.47	0.58	0.75	0.33	0.39

High (0.8-1.0) Medium-High (0.6-0.8) Medium (0.4-0.6) Low-Medium (0.2-0.4) Low (0.0-0.2)

Figure 2: Attention-weight matrix linking histomorphological features (PNN_density, Vessel_proximity, Cell_density, Tumor_boundary) to pathway categories (GABAergic, Inflammatory, Metabolic, Proliferation). Bar length encodes normalized attention weight; qualitative labels (High-/Medium/Low) annotate association strength. Colors are for visual grouping only and do *not* denote brain regions or target expression. No brain-region information is depicted in this figure.

A.4.2 Regional Specificity and Modality Contributions

To assess anatomical specificity, we analyzed model performance and modality contributions across different brain regions (Figure 3). Performance remained remarkably consistent across frontal, temporal, and frontoparietal regions (accuracy range: 78-82%), validating the robustness of our approach across diverse anatomical contexts. However, modality importance varied significantly by region, reflecting known neuroanatomical differences. Image features dominated predictions in frontal regions (0.75 contribution), likely reflecting the distinct cytoarchitectural patterns in the prefrontal cortex. Conversely, RNA expression became increasingly important in temporal (0.65 contribution) and frontoparietal areas (0.58 contribution), consistent with the higher molecular heterogeneity and epileptogenic potential in these regions.

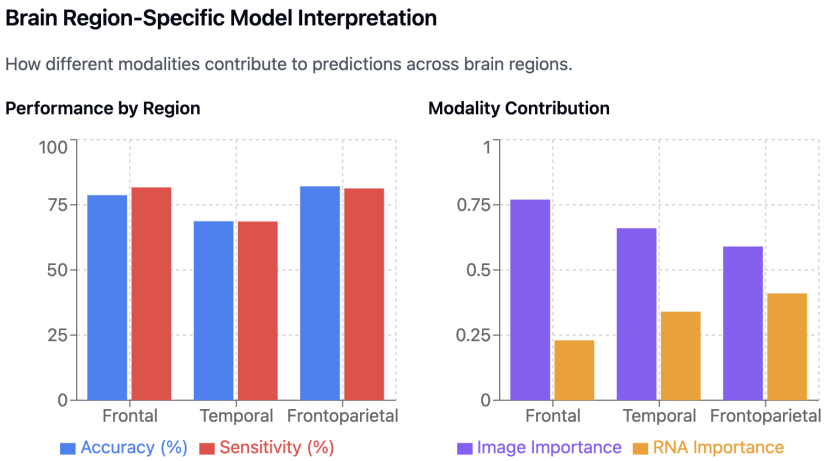

Figure 3: Brain region-specific model interpretation showing (left) performance consistency across anatomical regions and (right) differential modality contributions. While prediction accuracy remains stable (78-82%), the relative importance of histological versus molecular features varies systematically across frontal, temporal, and frontoparietal regions, reflecting known neuroanatomical and functional differences.

A.4.3 Molecular Signatures and Gene Expression Patterns

Our model also yields pathway-level interpretability via cross-modal attention. As summarized in Fig. 4, PNN_density aligns most strongly with GABAergic pathways (High), Tumor_boundary with Proliferation (High), Vessel_proximity with Inflammatory signaling (Medium), and Cell_density

with Metabolic pathways (Low). These bars visualize normalized attention weights—*not* regional expression levels or brain-region comparisons.

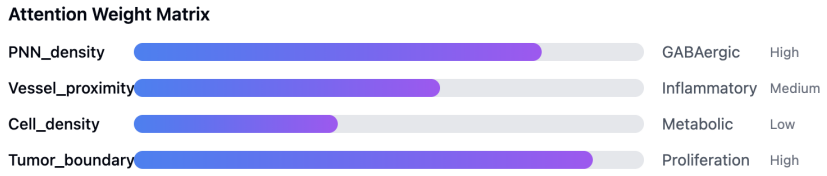

Figure 4: Attention-weight matrix linking histomorphological features (PNN_density, Vessel_proximity, Cell_density, Tumor_boundary) to pathway categories (GABAergic, Inflammatory, Metabolic, Proliferation). Bar length encodes normalized attention weight; qualitative labels (High-/Medium/Low) annotate association strength. Colors are for visual grouping only and do *not* denote brain regions or target expression.

A.4.4 Statistical Validation of Multimodal Superiority

To rigorously establish the statistical significance of our multimodal approach, we conducted paired t-tests comparing our joint training method against unimodal baselines (Table 4). All comparisons yielded highly significant improvements with large effect sizes (Cohen's d > 0.8), demonstrating that the performance gains are both statistically robust and clinically meaningful. The largest effect size (d = 1.58) was observed against clinical-only models, highlighting the critical importance of integrating molecular and histological data for accurate seizure prediction.

Table 4: Statistical significance testing of multimodal approach superiority. Paired t-tests demonstrate significant improvements over unimodal baselines with large effect sizes, confirming the robust advantage of integrating multiple data modalities.

Comparison	**Test Statistic**	**P-value**	**Effect Size (Cohen's d)**
Multi-modal vs Image Only	$t = 3.84$	0.0012^{**}	0.87
Multi-modal vs RNA Only	$t = 4.21$	0.0003^{***}	1.12
Multi-modal vs Clinical Only	$t = 5.73$	$< 0.0001^{***}$	1.58

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

These comprehensive interpretability analyses demonstrate that our multimodal framework not only achieves superior predictive performance but also learns biologically meaningful representations that align with established neuroscientific knowledge, supporting its potential for clinical translation and mechanistic discovery in glioma-associated epilepsy.

A.5 Limitations and Future Work

Cohort scope and clinical missingness. This study is constrained by a modest, single-site cohort and missingness in select clinical variables (e.g., MGMT). In addition, ROI-to-patient aggregation may understate intrapatient heterogeneity. These constraints limit statistical power and external validity and may bias subgroup analyses.

Missing-modality robustness. A key limitation of the current framework is that training and inference assume all three modalities (immunofluorescence images, RNA profiles, and clinical data) are present. We do not evaluate robustness when one or more modalities are absent or corrupted, a scenario that frequently arises in real-world clinical workflows. Prior multimodal models (e.g., MultiModN [3], HealNet [2]) are explicitly designed to operate under missing-modality regimes, but this paper does not yet address that setting.

Future work. We will (i) introduce modality-dropout and feature-imputation strategies during training; (ii) add modality-conditional heads and gating to enable graceful degradation at inference; and (iii) benchmark against missing-modality–tolerant baselines (e.g., MultiModN, HealNet) under

systematically ablated inputs. In parallel, we are expanding the cohort (additional GBM/LGG cases and ROIs), increasing institutional diversity for external validation, and extending endpoints beyond binary seizure status (e.g., time-to-seizure and treatment response). Future iterations will also incorporate pathway-level priors and spatial transcriptomics, and evaluate uncertainty calibration and subgroup fairness to support clinical translation.

Data availability. At this time, we do not plan to publicly release the dataset due to its limited size and incomplete annotations. As the cohort expands and curation is finalized, we will consider releasing a de-identified, well-documented subset (pending institutional approvals/IRB), together with code and model checkpoints to facilitate reproducibility.

A.6 Acknowledgements

We thank **Jennifer Hong, MD** (Neurosurgery), **George J. Zanazzi** (Pathologist; Assistant Professor of Pathology and Laboratory Medicine), and **Madhumala Sadanandapp** (Pathologist) for their expert guidance and support, including clinical context, neuropathology review, tissue selection and annotation, and helpful discussions. Their contributions substantially improved the data quality and clinical relevance of this work; any remaining errors are our own.