From Vision to Graph Self-Supervised Learning in Digital Pathology

Sevda Öğüt [*1] **Carlos Hurtado** [*] **Cédric Vincent-Cuaz** [1] **Natalia Dubljevic** **Vaishnavi Subramanian** [1]
Dorina Thanou [1] **Pascal Frossard** [1]

Abstract

Although vision-based self-supervised learning is revolutionizing digital pathology, its domain-agnostic architectures may fail to adequately focus on the primary biological components in tissues, namely the cells and their complex interactions. We therefore propose to transform tissues into biologically informed cell graphs and investigate the effectiveness of graph SSL in encoding them. We demonstrate that pre-training on a large collection of patches using GraphMAE, with heterophilic graph neural networks, yields on par performances against popular vision-based SSL models, while using significantly fewer parameters. Finally, we show that the learned graph embeddings can effectively complement their vision-based counterparts by using a late multi-modal fusion strategy.

1. Introduction

Recent advances in large-scale self-supervised learning have enabled the design of domain-specific vision models that are transformative for digital pathology (Wang et al., 2022; Filiot et al., 2023; Chen et al., 2024). These models analyze high-resolution pan-cancer whole-slide images (WSIs), with significant heterogeneity across different biological scales, to address clinically relevant tasks, such as cancer typing and grading, treatment response assessment, and survival prediction. Traditionally, they operate on small patches extracted from WSIs. These patches are encoded using vision transformers that model each patch as a set of smaller, non-overlapping patches called tokens (Dosovitskiy et al., 2020). While these tokens constitute the primary information units to be processed and aggregated by the architecture, they are misaligned with the core biological entities present in these patches, namely the cells (Shafi & Parwani, 2023; Chen et al., 2022).

[*]Equal contribution [1]LTS4, EPFL, Lausanne, Switzerland. Correspondence to: Sevda Öğüt <sevda.ogut@epfl.ch>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

In our work, we postulate that explicitly modeling the cells and their spatial arrangement could lead to an informative tissue encoding that improves upon or complements the implicit representations learned by vision models. To this end, we propose to model each patch as a cell graph (Pati et al., 2022), in which nodes represent cells described by biological features and edges reflect the spatial organization of the cells. We construct a pre-training dataset by representing 1126 H&E-stained breast cancer WSIs from the TCGA database (Weinstein et al., 2013) as cell graphs, resulting in 11 million patch-level graphs. We then learn a self-supervised representation of these graphs using the GraphMAE framework (Hou et al., 2022), a state-of-the-art masked autoencoder for graph data. Importantly, we propose to use heterophilic Graph Neural Networks (GNNs) for the encoding and decoding stages of GraphMAE as they are well adapted to the heterogeneity of tumor environments (Luan et al., 2022).

To ensure a fair comparison of the representational capabilities of our graph-based model, we perform an analogous pre-training scheme in the vision domain using the well-known DINOv2 (Oquab et al., 2023) and MAE (He et al., 2022; Hou et al., 2022) frameworks. Additionally, we assess the effect of combining graph- and vision-based representations through a late multi-modal fusion scheme (Jiao et al., 2024). The resultant embeddings of each model are finally evaluated on cancer typing and grading tasks, using the pre-training dataset and three additional downstream datasets, respectively. Our results demonstrate that graph-based representations are at least as discriminative as vision-based ones and that the two representations complement each other well, motivating future investigations into multi-modal pre-training strategies.

2. Proposed Approach

2.1. Cell-graph construction

We follow the preprocessing workflow outlined in Figure 1 on publicly available breast cancer datasets. First, we apply the method described in Campanella et al. (2019) to detect tissue regions from the H&E-stained WSIs at 20x magnification (i.e., 0.5 μm/pixel) and subsequently patch each image into non-overlapping 224×224 pixel tiles. Notice that this

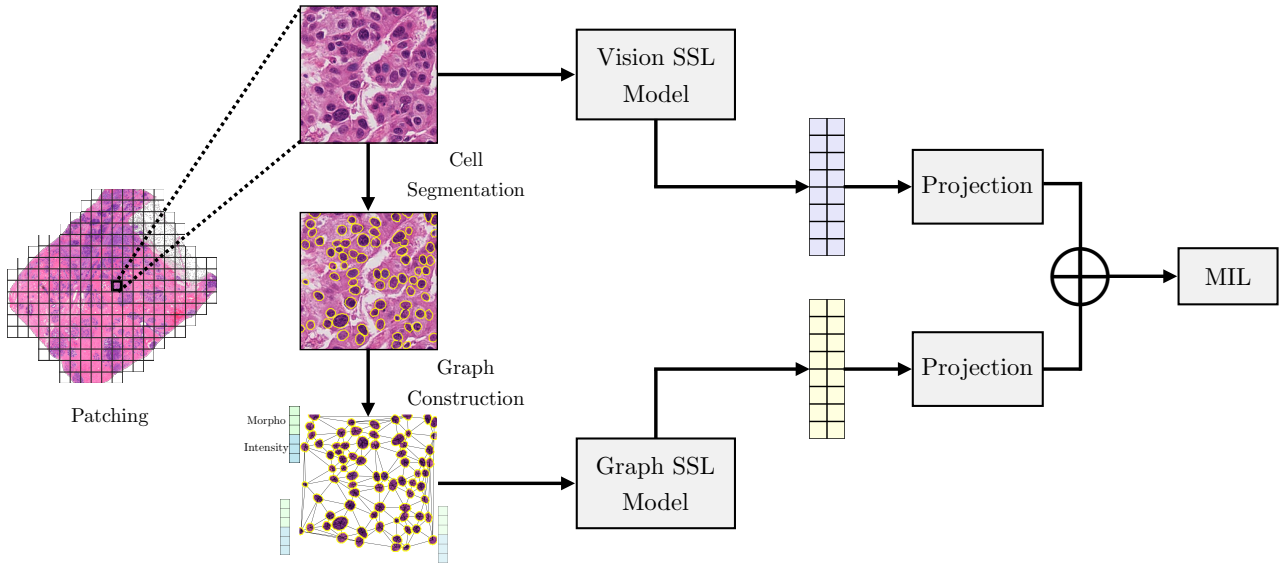

Figure 1. Pipeline showing cell-graph construction, vision- and graph-based self-supervised learning, and multi-modal late fusion with an example whole slide image from the TCGA BRCA dataset. The symbol $\bigoplus$ represents a concatenation operation.

choice of tile size is essentially made to align vision and graph models as detailed in the following sections. Next, we utilize a pre-trained cell segmentation model, StarDist 2D (Schmidt et al., 2018), to segment individual cells within these tiles. The cells are then considered as graph nodes. Inspired by Zhao et al. (2023) and Fournier et al. (2025), we assign 44 handcrafted features describing the cell morphology and color intensities to each node, which are known to be discriminant across cell types. These features are detailed in Table 4 of the Appendix. Lastly, we model the relative spatial arrangement of cells using Delaunay triangulation, mimicking an adaptive nearest neighbor graph. This results in a sparse set of edges weighted by the distance in μm between connected cells.

Following the scheme described above, we construct a pre-training graph dataset derived from 1126 H&E-stained breast cancer WSIs from the TCGA database (Weinstein et al., 2013), resulting in 11 million cell graphs.

2.2. Graph self-supervised learning

For learning graph representations, we adopt the generative self-supervised learning (SSL) framework GraphMAE (Hou et al., 2022), which has been shown to outperform most contrastive SSL methods predominant in the field. While more recent variants exist (Hou et al., 2023; Tan et al., 2023; Wang et al., 2024; Bai et al., 2024), we emphasize that GraphMAE still provides one of the best balances of computational efficiency and performance. It adapts the masked-autoencoding paradigm to graph-structured data by randomly dropping a fraction of nodes' features from the input graph while preserving the graph topology and using a GNN as the encoder. The previously masked nodes are then re-masked in the decoding stage, and their initial features are reconstructed using another 1-layer GNN.

In contrast to the homophilic GNNs used in GraphMAE for encoding and decoding, we propose to use heterophilic GNNs, which are better suited to the inherent heterogeneity of tumor microenvironments. Specifically, we adopt the Adaptive Channel Mixing architecture, with filters derived from the renormalized random walk matrices (Luan et al., 2022; Xu et al., 2018a). Additionally, we incorporate a Jumping Knowledge strategy, acting as residual connections, by concatenating outputs of each layer before projecting them onto a desired dimension with a projection head (Xu et al., 2018b). Our compact architecture is designed to effectively encode the cell graphs while maintaining a linear complexity in the number of cells.

2.3. Multi-modal fusion of images and graphs

While our graph framework provides a flexible means of incorporating prior biological knowledge, the remarkable performance of recent vision-based approaches strongly supports their ability to capture discriminant tissue properties (Chen et al., 2024). We hypothesize that graph models could provide complementary information to vision models to further improve their performance. Therefore we investigate

the benefits of graph and vision multi-modal fusion considering two types of vision-based models independently, namely DINOv2 (Oquab et al., 2023) and MAE (He et al., 2022). In practice, DINOv2, which builds on self-distillation using a student-teacher pair of vision transformers, is nowadays the most studied and utilized approach (Campanella et al., 2024). However, previously introduced approaches, including masked autoencoding frameworks like MAE, remain competitive and are preferred for generative tasks (Kraus et al., 2024).

For patch-level or WSI-level downstream tasks, we propose to follow a late fusion scheme for each patch (Jiao et al., 2024), which consists of concatenating image and graph embeddings after mapping each of them to a lower-dimensional space separately. These embeddings are respectively given by the CLS token of the vision transformer encoder and the mean of node embeddings from the GNN encoder, as mean pooling provides a simple yet effective method for obtaining a graph-level representation.

3. Experimental Design

3.1. Datasets

In the following, we evaluate the model embeddings on four slide-level tasks. The first task corresponds to our pre-training dataset, referred to as the in-domain dataset. The remaining three tasks use out-of-domain (OOD) region-of-interest (RoI) datasets, namely BACH, BRACS, and BreakHis, to assess the transferability of the learned representations across diverse image acquisition protocols, data sources, and tissue types.

3.2. Benchmark settings

All SSL models discussed in Section 2 were pre-trained for 100 epochs using the optimization strategies described in their original papers, fixing the batch size to 2048. For each method, we select the epoch that minimizes the validation loss. For vision models, we use the standard hyperparameters provided by the authors, while using ViT-B/16 as the backbone. For GraphMAE, we test different embedding dimensions in $\{512, 768, 1024\}$, masking ratios in $\{0.50, 0.75\}$, and replacement ratios in $\{0.00, 0.10\}$, following the guidelines from Hou et al. (2022). We set the number of GNN layers to five in the graph encoder, which corresponds to the averaged graph diameter in the pre-training dataset. Patch-level embeddings are then extracted from each model and for each evaluation dataset, following the methodology described in Section 2.3.

Since the final tasks are either at the WSI or the RoI level, we leverage multiple instance learning (MIL) methods to aggregate the patch-level embeddings into a single slide-level embedding. We benchmark three state-of-the-art attention-based MIL approaches, namely ABMIL (Ilse et al., 2018), add-ABMIL (Javed et al., 2022), and conj-ABMIL (Early et al., 2024). They mainly differ in how they couple the attention mechanism with the final two-layer MLP classifier. Finally, for each of the SSL and MIL methods, we perform a 5-fold CV to validate their hyperparameters, relying on macro F1 scores on validation sets.

Table 1. Mean ± standard deviation (%) of test macro F1 scores for benchmarked unimodal SSL methods. In-domain setting refers to the result from TCGA BRCA and out-of-domain setting refers to the averaged results from BACH, BRACS, and BreakHis. Best results are bolded and second-best results are underlined for each MIL method.

	DINOv2	MAE	GraphMAE
In-Domain			
ABMIL	60.14 ± 3.09	**78.07 ± 1.44**	<u>68.60 ± 0.84</u>
add-ABMIL	53.33 ± 9.00	<u>56.30 ± 10.32</u>	**64.23 ± 2.06**
conj-ABMIL	62.30 ± 1.03	**77.86 ± 1.83**	<u>70.96 ± 3.09</u>
Out-of-Domain			
ABMIL	60.01 ± 0.82	**75.66 ± 1.17**	<u>74.09 ± 1.74</u>
add-ABMIL	60.79 ± 2.28	<u>73.03 ± 2.55</u>	**75.22 ± 2.22**
conj-ABMIL	58.05 ± 2.11	**76.17 ± 1.02**	<u>74.37 ± 1.66</u>

4. Results

4.1. Comparison of vision and graph SSL

We first compare the global embedding quality of each unimodal SSL method by reporting their aggregated downstream performances across MIL frameworks on both in- and out-of-domain datasets in Table 1. Results per dataset can be found in Table 6 in the Appendix. Note that we selected the GraphMAE model leading to the best averaged performances across MIL methods for this analysis. We observe that GraphMAE embeddings perform comparably with those of MAE on our OOD benchmark, despite having about 12 times fewer parameters (see details in Appendix, Table 5). This highlights the value of explicitly encoding biological priors through our graph-based modeling.

However, we can see that GraphMAE lags behind MAE on in-domain performance across most MIL methods. To explain this phenomenon, we investigated via Principal Component Analysis whether each method was fully exploiting its available dimensions in the embedding space to represent each dataset. Results reported in Figure 3 in the Appendix show that our graph models suffer from a drastic dimension collapse issue on the pre-training dataset. Moreover, a similar problem is observed for the OOD dataset BRACS, but not for the other two, BACH and BreakHis, where GraphMAE is most competitive against vision models. However, no

Table 2. Absolute gains over unimodal vision models of our multi-modal strategy. In-domain setting refers to the result from TCGA BRCA and out-of-domain setting refers to the averaged results over BACH, BRACS, and BreakHis.

	GraphMAE+DINOv2	GraphMAE+MAE
In-Domain		
ABMIL	+4.49	+0.54
add-ABMIL	-8.2	+20.43
conj-ABMIL	+1.74	-0.68
Out-of-Domain		
ABMIL	+10.60	+4.54
add-ABMIL	+11.65	+7.33
conj-ABMIL	+13.32	+3.65

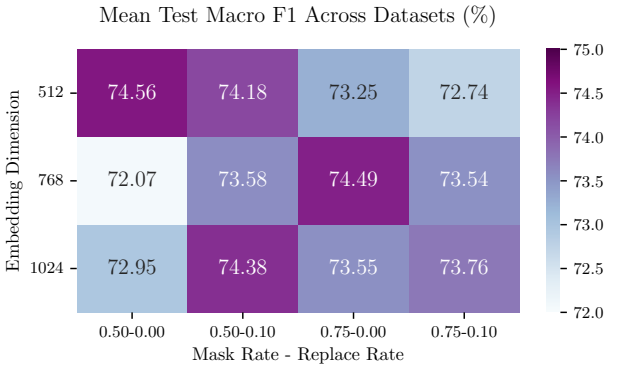

Figure 2. Averaged test macro F1 scores across datasets of various hyperparameter configurations of the GraphMAE model.

significant dimension collapse is observed for vision models, which rely on transformers and thus on self-attention. These differences between domains may indicate a lack of relational encoding capabilities in the GNN architectures we have chosen, the improvement of which could be an interesting avenue for future work.

Finally, we note that GraphMAE and MAE embeddings perform significantly better than DINOv2 in both in- and out-of-domain contexts. This may be due to our relatively smaller scale of data compared to the scales where DINOv2 performed remarkably well (Oquab et al., 2023).

4.2. Multi-modal embeddings

Next, we investigate whether the graph embeddings introduced above can effectively complement the vision embeddings within the multi-modal framework described in Section 2.3. To this end, we report the absolute gains of our multi-modal strategy over unimodal vision models in Table 2. In the OOD setting, fusing DINOv2 or MAE with GraphMAE consistently yields significant performance improvements. While gains are also observed in the in-domain evaluation, they are more marginal, likely due to the dimension collapse issue noted earlier. An exception is add-ABMIL, which exhibits substantial instability on the pre-training dataset. Overall, these results underscore how image- and graph-based self-supervised embeddings can capture complementary information.

4.3. Sensitivity analysis on graph models

To further assess the robustness of our graph-based methodology, we conduct a sensitivity analysis on its validated hyperparameters, detailed in Section 3.2. We report the averaged test macro F1 scores over the OOD datasets and MIL frameworks in Figure 2. Interestingly, we observe that GraphMAE manages to produce similar discriminant representations for all embedding dimensions. Naturally, if the dimension is higher, this performance is obtained for higher masking and replacement rates, which respectively induce implicit regularization. Besides, we observe that performances are relatively stable across all tested hyperparameters, with only a 2.49% difference between the best and the worst configuration, which confirms the robustness of our approach.

5. Conclusion

In this work, we show that modeling digital pathology patches as cell graphs, before encoding them via an adapted SSL approach, leads to comparable performance with well-known vision models while using far fewer parameters. In addition, we demonstrate that a simple late fusion strategy can effectively exploit graph embeddings to complement vision-based ones.

Nevertheless, our graph-based approach is significantly prone to dimension collapse in some scenarios, which likely limits its performance. To remedy this problem, we envisage three main avenues that need to be explored further. First, we plan to improve the relational encoding capabilities of our sparse GNNs, for instance via graph transformers (Yuan et al., 2025), while keeping in mind that such architectures may lead to additional challenges in terms of scalability. Second, we consider adding explicit regularizations to the GraphMAE model. Third, we plan to further exploit the flexibility of our graph framework by augmenting the set of chosen node features, e.g., including texture features or even more advanced single-cell data, as well as potentially adding additional edge feature information relating to the extracellular matrix. We believe addressing these challenges and further investigating multi-modal fusion strategies between vision and graph models, will soon guide the development of better foundation models for digital pathology.

Impact Statement

This work advances computational pathology by proposing a novel graph-based self-supervised learning approach that explicitly models cellular arrangements, complementing traditional vision-based models. By better aligning learned representations with biological structures, this research contributes to improved interpretability and robustness in breast cancer characterization tasks. Although the immediate clinical impact is limited due to the scale of generated graphs, the proposed methodology lays the foundation for future large-scale clinical applications. Ethical considerations related to patient privacy from WSIs are addressed through rigorous anonymization and secure data management protocols, effectively mitigating risks associated with patient data re-identification.

Acknowledgements

The authors would like to thank Julian David Jülg for his valuable feedback and acknowledge the Research Computing Platform (RCP) at EPFL for providing computational resources. Carlos Hurtado gratefully acknowledges the support of a fellowship from the la Caixa Foundation (ID 100010434), under fellowship code LCF/BQ/EU24/12060071.

References

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., et al. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56: 122–139, 2019.

Bai, L., Xu, Z., Cui, L., Li, M., Wang, Y., and Hancock, E. Hc-gae: The hierarchical cluster-based graph autoencoder for graph representation learning. *NeurIPS*, 37: 127968–127986, 2024.

Brancati, N., Anniciello, A. M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8): 1301–1309, 2019.

Campanella, G., Vanderbilt, C., and Fuchs, T. Computational pathology at health system scale – self-supervised foundation models from billions of images. In *AAAI Spring Symposium on Clinical Foundation Models*, 2024.

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., and Mahmood, F. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *IEEE CVPR*, pp. 16144–16155, 2022.

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., Song, A. H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

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 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Early, J., Cheung, G., Cutajar, K., Xie, H., Kandola, J., and Twomey, N. Inherently interpretable time series classification via multiple instance learning. In *ICLR*, 2024.

Filiot, A., Ghermi, R., Olivier, A., Jacob, P., Fidon, L., Kain, A. M., Saillard, C., and Schiratti, J.-B. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757.

Fournier, L., Haefliger, G., Vernhes, A., Jung, V., Letovanec, I., Frossard, P., Vincent-Cuaz, C., and Luisier, R. Extended pre-training of histopathology foundation models uncovers co-existing breast cancer archetypes characterized by rna splicing or tgf-β dysregulation. *bioRxiv*, 2025. doi: 10.1101/2025.03.27.645192.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *IEEE CVPR*, pp. 16000–16009, 2022.

Hou, Z., Liu, X., Cen, Y., Dong, Y., Yang, H., Wang, C., and Tang, J. Graphmae: Self-supervised masked graph autoencoders. In *ACM SIGKDD*, pp. 594–604, 2022.

Hou, Z., He, Y., Cen, Y., Liu, X., Dong, Y., Kharlamov, E., and Tang, J. Graphmae2: A decoding-enhanced masked self-supervised graph learner. In *ACM Web Conference*, pp. 737–746, 2023.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *ICML*, 2018.

Javed, S. A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., and Prakash, A. Additive mil: Intrinsically interpretable multiple instance learning for pathology. In *NeurIPS*, 2022.

Jiao, T., Guo, C., Feng, X., Chen, Y., and Song, J. A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua*, 80(1), 2024.

Kraus, O., Kenyon-Dean, K., Saberian, S., Fallah, M., McLean, P., Leung, J., Sharma, V., Khan, A., Balakrishnan, J., Celik, S., et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *IEEE CVPR*, pp. 11757–11768, 2024.

Luan, S., Hua, C., Lu, Q., Zhu, J., Zhao, M., Zhang, S., Chang, X.-W., and Precup, D. Revisiting heterophily for graph neural networks. In *NeurIPS*, volume 35, pp. 1362–1375, 2022.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Pati, P., Jaume, G., Foncubierta-Rodriguez, A., Feroce, F., Anniciello, A. M., Scognamiglio, G., Brancati, N., Fiche, M., Dubruc, E., Riccio, D., et al. Hierarchical graph representations in digital pathology. *Medical Image Analysis*, 75:102264, 2022.

Schmidt, U., Weigert, M., Broaddus, C., and Myers, G. Cell detection with star-convex polygons. In *MICCAI*, pp. 265–273, 2018. doi: 10.1007/978-3-030-00934-2_30.

Shafi, S. and Parwani, A. V. Artificial intelligence in diagnostic pathology. *Diagnostic Pathology*, 18(1):109, 2023.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *IEEE Biomedical Engineering*, 63(7):1455–1462, 2015.

Tan, Q., Liu, N., Huang, X., Choi, S.-H., Li, L., Chen, R., and Hu, X. S2gae: Self-supervised graph autoencoders are generalizable learners with graph masking. In *ACM WSDM*, pp. 787–795, 2023.

Vadori, V., Peruffo, A., Graïe, J.-M., Finos, L., and Grisan, E. Automated classification of cell shapes: A comparative evaluation of shape descriptors. In *IEEE ISBI*, pp. 1–4. IEEE, 2025.

Wang, L., Tao, X., Liu, Q., and Wu, S. Rethinking graph masked autoencoders through alignment and uniformity. In *AAAI*, volume 38, pp. 15528–15536, 2024.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018a.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. Representation learning on graphs with jumping knowledge networks. In *ICML*, pp. 5453–5462, 2018b.

Yuan, C., Zhao, K., Kuruoglu, E. E., Wang, L., Xu, T., Huang, W., Zhao, D., Cheng, H., and Rong, Y. A survey of graph transformers: Architectures, theories and applications. *arXiv preprint arXiv:2502.16533*, 2025.

Zhang, D., Lu, G., et al. A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *ICIMADE*, pp. 1–9, 2001.

Zhao, S., Chen, D.-P., Fu, T., Yang, J.-C., Ma, D., Zhu, X.-Z., Wang, X.-X., Jiao, Y.-P., Jin, X., Xiao, Y., et al. Single-cell morphological and topological atlas reveals the ecosystem diversity of human breast cancer. *Nature Communications*, 14(1):6796, 2023.

Appendix

5.1. Dataset details

TCGA BRCA. We leverage this breast invasive carcinoma dataset, which comprises 1126 H&E-stained WSIs, sized up to 100K×100K pixels. We obtain approximately 11 million patches from this dataset that serve two purposes in our study: (1) self-supervised pre-training to learn vision and graph representations, and (2) in-domain downstream evaluation corresponding to the classification of infiltrating ductal carcinoma (794) and lobular carcinoma (204) of different grades. Note that multiple other subtypes were excluded in this study due to their small sample size.

BACH. We use this breast cancer histology dataset (Aresta et al., 2019) of 400 region-of-interest (RoI) images for out-of-domain downstream evaluation. The pre-trained model embeddings are fine-tuned for a 4-class classification task of cancer subtyping distributed as normal (100), benign (100), in situ carcinoma (100), and invasive carcinoma (100).

BRACS. Similarly, we use this breast cancer dataset put forth by Brancati et al. (2022) of 4539 labeled RoIs for fine-tuning the pre-trained model embeddings and downstream evaluation. The end task consists of a 7-class classification of tumor subtypes, distributed as normal (484), pathological benign (836), usual ductal hyperplasia (517), flat epithelial atypia (756), atypical ductal hyperplasia (507), ductal carcinoma in situ (790), and invasive carcinoma (649).

BreakHis. The final out-of-domain setting contains a dataset of 1995 labeled microscopic images of breast tumor tissue collected from 82 patients (Spanhol et al., 2015). The downstream task is again an 8-class classification of tumor subtype, with the labels of adenosisc (114), fibroadenoma (253), phyllodes tumor (109), tubular adenoma (149), ductal carcinoma (864), lobular carcinoma (156), mucinous carcinoma (205), and papillary carcinoma (145).

Table 3. Summary statistics for the in- and out-of-domain datasets.

Dataset	# Slides	Avg # Nodes	Avg # Edges
TCGA-BRCA	1126	43.94	119.16
BACH	400	35.05	92.93
BRACS	4539	46.96	128.09
BreakHis	1995	17.26	41.43

5.2. Hand-crafted features

We use the features shown in Table 4 as node attributes for our cell graphs. Please refer to (Zhang et al., 2001) for a detailed explanation of Fourier features with centroid signature and to (Vadori et al., 2025) for their use as cell shape descriptors.

Table 4. The set of cell-level shape and intensity descriptors used as input node features in our graph construction pipeline. R, G, and B stand for red, green, and blue, respectively.

Feature Names	
min intensity R	mean intensity R
min intensity G	mean intensity G
min intensity B	mean intensity B
max intensity R	var intensity R
max intensity G	var intensity G
max intensity B	var intensity B
probability	eccentricity
orientation	area
axis major length	perimeter
axis minor length	Fourier features

5.3. Implementation details

All experiments were run on NVIDIA A100 (80GB) and H200 (140GB) GPUs and the number of parameters that require gradients in DINOv2, MAE, and GraphMAE models can be found in the Table 5.

Table 5. Model size comparison across SSL methods. We denote by d the validated embedding dimension for GraphMAE models.

Model	**Number of Parameters** (10^6)
DINOv2	171.63
MAE	111.91
GraphMAE ($d = 512$)	9.34
GraphMAE ($d = 768$)	20.88
GraphMAE ($d = 1024$)	37.02

5.4. Additional results

Table 6 contains per-dataset results for all unimodal and multi-modal architectures and Figure 3 downstrates the explained variance ratios of the principal components in all four datasets.

Table 6. Mean ± standard deviation (%) of test macro F1 scores over 5 runs for TCGA-BRCA, BACH, BRACS, and BreakHis for the models of DINOv2, MAE, GraphMAE, and the multi-modal late fusion combinations.

	DINOv2	MAE	GraphMAE	DINOv2+GraphMAE	MAE+GraphMAE	DINOv2+MAE
TCGA-BRCA						
ABMIL	60.14 ± 3.09	78.07 ± 1.44	68.60 ± 0.84	64.63 ± 2.03	78.61 ± 1.96	77.08 ± 0.87
add-ABMIL	53.33 ± 9.00	56.30 ± 10.32	64.23 ± 2.06	45.13 ± 16.70	76.73 ± 1.61	75.70 ± 2.09
conj-ABMIL	62.30 ± 1.03	77.86 ± 1.83	70.96 ± 3.09	64.04 ± 3.32	77.18 ± 1.37	77.43 ± 1.52
BACH						
ABMIL	56.02 ± 1.38	60.34 ± 2.42	68.56 ± 2.74	64.93 ± 2.80	71.78 ± 4.13	77.53 ± 2.37
add-ABMIL	56.81 ± 3.87	52.03 ± 5.41	70.70 ± 3.65	68.07 ± 2.10	71.82 ± 3.05	71.80 ± 4.32
conj-ABMIL	52.92 ± 2.87	61.97 ± 1.30	69.51 ± 3.20	67.05 ± 2.43	70.42 ± 2.26	76.68 ± 2.22
BRACS						
ABMIL	44.95 ± 0.62	69.73 ± 0.35	61.90 ± 1.14	57.02 ± 1.76	71.53 ± 1.15	68.45 ± 0.78
add-ABMIL	48.04 ± 0.58	70.34 ± 0.93	63.13 ± 0.59	57.38 ± 0.86	71.38 ± 0.89	66.73 ± 0.70
conj-ABMIL	43.70 ± 0.55	70.03 ± 0.60	61.23 ± 0.34	55.97 ± 0.91	71.55 ± 0.64	68.00 ± 0.70
BreakHis						
ABMIL	79.06 ± 0.45	96.90 ± 0.73	91.81 ± 1.34	89.88 ± 2.09	97.29 ± 0.72	98.06 ± 0.87
add-ABMIL	77.53 ± 2.38	96.71 ± 1.30	91.82 ± 2.41	91.88 ± 0.47	97.87 ± 0.72	98.64 ± 0.47
conj-ABMIL	77.52 ± 2.91	96.52 ± 1.15	92.38 ± 1.44	91.08 ± 1.13	97.49 ± 1.44	97.67 ± 1.00

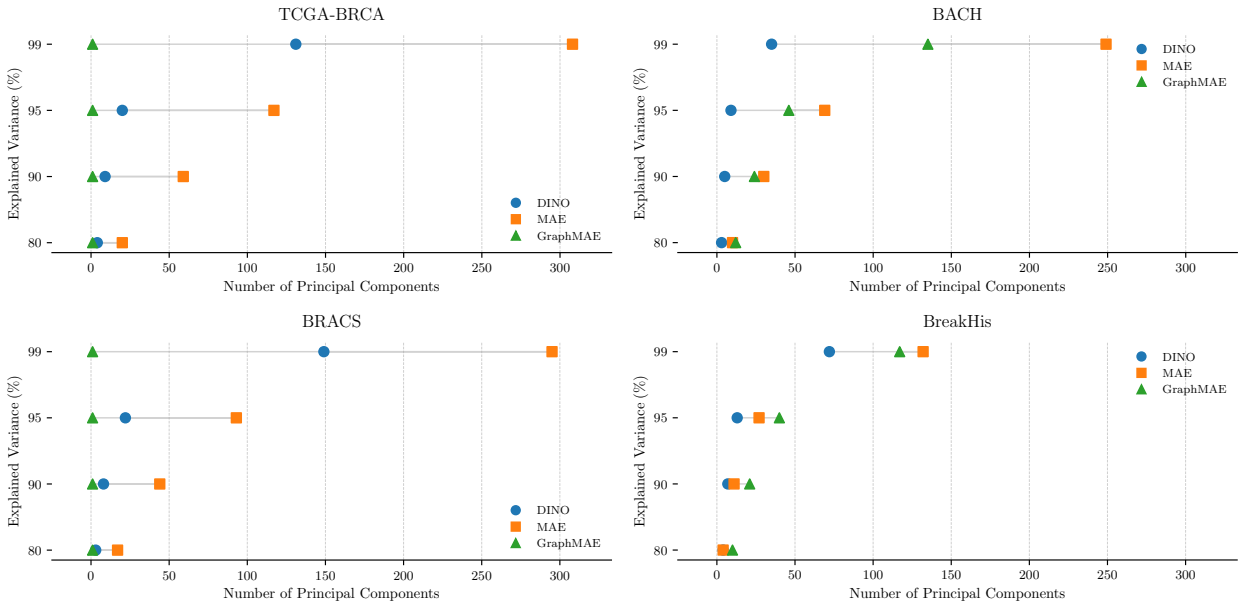

Figure 3. Number of principal components required by DINOv2, MAE, and GraphMAE to reach various explained-variance thresholds across the four datasets.