Probing, Fusion, and Trustworthiness: A Systematic Evaluation of Foundation Model Representations for Multimodal Cancer Analysis

Jingyu Hu [1 2] **Giuseppe Tripodi** [1 3] **Reed Naidoo** [1 4] **Sarah F McGough** [5] **Tapabrata Chakraborti** [1]

Abstract

Foundation models (FMs) have emerged as powerful representation extractors for medical data, yet their generalizability to datasets under distribution shift remains underexplored. This work systematically evaluates FM-based representations on a suite of computational pathology tasks across two real-world commercial cohorts, IH-BC and IH-NSCLC, drawn from the licensed in-house (IH) oncology dataset. The analysis focuses on two modalities, whole-slide images and transcriptomic profiles, drawn from the IH multimodal data. We first benchmark unimodal probing performance across five FMs on eight downstream classification tasks, and find that image and omics representations carry complementary predictive signals. Then we investigate whether multimodal fusion can yield additional gains over unimodal baselines by comparing three image-omics fusion strategies built on paired representations. The trustworthiness of selected unimodal and multimodal pipelines is further assessed through conformal prediction. Our results show that FM representations achieve competitive performance on out-of-distribution data and that multimodal fusion helps mainly when no single modality dominates the signal. Conformal prediction reveals that in the majority of cases where a point prediction fails, the true diagnosis remains recoverable within the prediction set, reinforcing the value of uncertainty-aware inference for clinical support.

[1]The Alan Turing Institute, London, United Kingdom [2]University of Bristol, Bristol, United Kingdom [3]University of Manchester, Manchester, United Kingdom [4]The Institute of Cancer Research, London, United Kingdom [5]Genentech, United States. Correspondence to: Tapabrata Chakraborty <tchakraborty@turing.ac.uk>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Artificial intelligence has shown promising performance across a range of cancer diagnosis applications, from medical imaging (Ardila et al., 2019; McKinney et al., 2020) and pathology (Campanella et al., 2019) to molecular and genomic interpretation (Capper et al., 2018; Jiao et al., 2020) and clinical outcome prediction (Placido et al., 2023). A common workflow in these methods applies a modality-specific encoder to extract data representations, which are then consumed by either a unimodal or multimodal learning module for downstream prediction. The expressiveness of the encoded representation largely determines the quality of the downstream predictor, and prior work has extensively studied how encoder architecture and capacity match each data modality (Geng et al., 2022; Zhang et al., 2020).

As medical datasets continue to grow in number and scale (Deng et al., 2026), training a dedicated encoder from scratch for every new task becomes computationally expensive. To address this bottleneck, recent work starts to explore medical Foundation Models (FMs), which are pretrained on large medical corpora and used as training-free feature extractors across heterogeneous downstream tasks. Both pathology FMs like CONCH (Lu et al., 2024a) and transcriptomic FMs like UCE (Rosen et al., 2023) have reported promising transferability, though evaluations are mainly performed on public benchmarks that can overlap with or are close to the pretraining corpora of FMs.

In real-world scenarios, however, there are also datasets collected from industrial and commercial sources that, due to different data collection pipelines, follow different distributions from the public ones. Whether FMs can generalize to such unseen data remains an underexplored question. This gap motivates our first question: *whether FM representations can transfer to OOD datasets and yield reliable representations under probing.* We study this question using an in-house (IH) real-world dataset of multi-cancer cases with paired whole-slide H&E images and transcriptomic profiles. We probe image representations from four image FMs and omics representations from three transcriptomic encoders and evaluate their performance on eight downstream classification tasks.

The evaluation on unimodal probing shows that image and

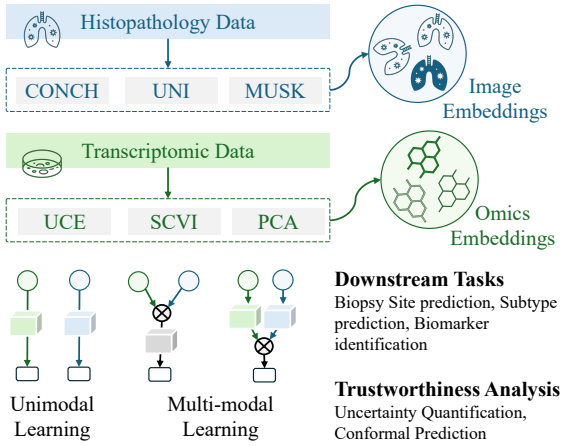

Figure 1. Overview of the Experimental Setups.

omics representations carry complementary signals. This is consistent with the real-world setting where prognosis and treatment response are jointly determined by multiple modalities such as morphology, molecular state, and clinical context. Prior work has therefore started multimodal learning exploration that applies fusion strategies such as concatenation (Chen et al., 2020) and cross-modal attention (Chen et al., 2021) to incorporate information across modalities. We therefore ask *whether image-omics fused representations can obtain additional predictive performance over the unimodal representations.* By pairing image and omics representations with three fusion strategies, we find multimodal fusion delivers stronger performance on some tasks.

The evaluations so far focus on predictive performance (e.g., accuracy). However, in high-stakes domains like medical diagnosis, high utility alone does not guarantee trustworthiness: a model may still be miscalibrated or unfair across demographic subgroups. This leads us to further wonder *how trustworthy are the unimodal and multimodal techniques studied above?* We apply conformal prediction (Angelopoulos & Bates, 2021) to assess predictive uncertainty and quantify the gap in predictive performance across subgroups. The selected pipelines achieve meaningful coverage guarantees and broadly similar performance across groups, while still revealing task-specific disparities for future work.

Figure 1 illustrates the overall workflow of our work, which proceeds in three parts: unimodal probing of image and omics representations across eight downstream tasks (§3), multimodal learning comparisons of three image-omics fusion strategies built on paired image-omics representations (§4), and trustworthiness evaluation of selected unimodal and multimodal techniques via conformal analysis (§5).

2. Experimental Setup

Datasets Preprocessing Methods. This study used the US-based deidentified Flatiron Health-Caris Life Sciences breast cancer and non-small cell lung cancer Clinical-Molecular Database (CMDB). Clinical data from the Flatiron Health Research Database (Flatiron Health, 2025; Ma et al., 2020; Zhang et al., 2021) are linked to molecular data, derived from Caris Life Sciences' MI Profile™ comprehensive profiling in the CMDB by probabilistic matching, providing a deidentified dataset (Caris Life Sciences, b;a;c). Based on cancer type, the collection is split into two in-house cohorts, referred to as IH-BC and IH-NSCLC. Each IH subset contains multimodal genomic data for the corresponding cancer. This study mainly takes two modalities of H&E images and omics data as input, and information about cancer subtypes, biopsy sites, and biomarkers is used as the downstream prediction tasks. Specifically, IH-BC includes LOH, Biomarker PR status, PIK3CA status, Biopsy Site, and Breast Cancer Subtype identification tasks, while IH-NSCLC includes Biopsy Site, Tumor Site and TMB identification tasks. The datasets are split into training, calibration, validation, and test sets at a ratio of 7:3:1:1. The multimodal learning methods are trained on the training set and evaluated on the test set. More detailed preprocessing steps are reported in Appendix B.1.

Modelling Methods. The experiment consists of two stages: representation extraction and representation learning. For representation extraction, we evaluated four tile-level foundation models (CONCH (Lu et al., 2024a), UNI (Chen et al., 2024), Virchow (Vorontsov et al., 2023) and MUSK (Xiang et al., 2025)) for H&E WSI data and one omics foundation model (UCE (Rosen et al., 2023)) for omics data. We also obtained omics representations using scVI (Lopez et al., 2018) and Principal Component Analysis (PCA) (Abdi & Williams, 2010) for comparison. In the representation learning stage, we considered five methods, each combined with different representation backbones, including two unimodal methods: H&E image-based Multiple Instance Learning (HEMIL) and Multilayer Perceptron for omics data (GeneMLP), and three multimodal methods: concatenation-based fusion (CONTACT), Multimodal Co-Attention Transformer (Chen et al., 2021) (MCAT), and Late Fusion Multiple Instance Learning (Naidoo et al., 2024) (LateMIL). All methods are designed for the same set of tasks, and further method details are provided in Appendix B.2.

Evaluation Metrics. Accuracy (ACC) and AUC are used as the main evaluation metrics for model utility performance of downstream classification tasks. The ROC curve is applied to illustrate the trade-off between the true positive rate and false positive rate at different classification thresholds. More detailed formulas can be found in Appendix B.3.

3. Experiments on Unimodal Probing

Table 1 summarizes the unimodal probing performance of four types of tiles representations and three omics represen-

Table 1. Unimodal Probing Performance on IH-BC and IH-NSCLC Tasks. We report both ACC and AUC.

		HEMIL				GeneMLP		
Dataset	Task	CONCH	UNI	MUSK	Virchow	UCE	SCVI	PCA
BC	Biopsy Site	0.8780 / 0.9111	0.8814 / 0.9343	0.8508 / 0.9191	0.8915 / 0.9412	0.6136 / 0.7000	0.7119 / 0.7590	0.7797 / 0.8467
	LOH	0.6566 / 0.7414	0.6128 / 0.7330	0.6431 / 0.7420	0.6229 / 0.7240	0.5253 / 0.6419	0.4882 / 0.7287	0.6667 / 0.7794
	PR	0.6933 / 0.7288	0.7133 / 0.7139	0.6900 / 0.7453	0.7267 / 0.7285	0.6200 / 0.6297	0.7333 / 0.7633	0.7700 / 0.8159
	PIK3CA	0.7500 / 0.7355	0.7433 / 0.7299	0.7533 / 0.6969	0.7500 / 0.7086	0.6967 / 0.6097	0.7033 / 0.7718	0.7933 / 0.7921
	Subtype	0.5824 / 0.8013	0.5714 / 0.8159	0.5788 / 0.7986	0.5678 / 0.7966	0.5311 / 0.7558	0.6520 / 0.8649	0.7106 / 0.8955
NSCLC	Biopsy Site	0.4602 / 0.6379	0.5057 / 0.6478	0.4830 / 0.6475	0.4716 / 0.6325	0.4375 / 0.5053	0.3295 / 0.5205	0.3977 / 0.5594
	TMB	0.6433 / 0.6691	0.6833 / 0.7081	0.6633 / 0.6210	0.6500 / 0.6769	0.6433 / 0.6364	0.5967 / 0.6542	0.6967 / 0.7277
	Tumor Site	0.4982 / 0.6283	0.4912 / 0.6520	0.4841 / 0.6177	0.4982 / 0.6575	0.4664 / 0.6100	0.4629 / 0.6014	0.4346 / 0.5915

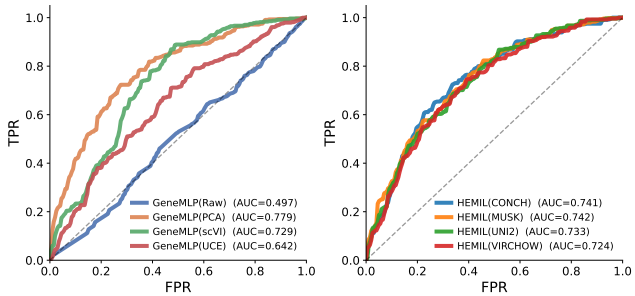

Figure 2. ROC Comparison on BC-LOH Task.

tations across eight downstream tasks.

Image foundation models achieve broadly comparable performance, though results vary considerably across cancer types: AUC exceeds 0.9 on IH-BC Biopsy Site across all image FMs, whereas IH-NSCLC Biopsy Site AUC ranges only from 0.6325 to 0.6478. Among tile representations, differences between image FMs are small relative to differences across tasks, indicating that downstream task difficulty is the key factor shaping unimodal probing performance. Omics representations provide strong unimodal signal on several tasks, and in some cases outperform image-based methods. For example, the highest accuracy on the task of BC-PIK3CA among image-based approaches is 0.7533, whereas the best omics-based result reaches 0.7933. The classical PCA baseline generally outperforms learned transcriptomic encoders on most task and achieves highest AUC on LOH (0.7794), PR (0.8159), PIK3CA (0.7921), Subtype (0.8955), and TMB (0.7277).

We also compared ROC curves (Figure 2) and observed a consistent pattern: image representations show more stable performance with relatively small variance across FMs. All three omics representation methods (UCE, PCA, and scVI) outperform direct modeling on the raw full gene expression profile on BC-LOH. However, the foundation model UCE underperforms the non-foundation-model approaches scVI and PCA. The pattern observed in the industrial IH dataset aligns with prior discussions on public datasets, suggesting that PCA can be better suited for capturing biological perturbations than existing omics foundation models (Bendidi et al., 2024). This indicates that building effective transcriptomic foundation models remains an open challenge.

4. Experiments on Multimodal Fusion

The above results show that omics and image modalities each have strengths on different tasks, we next investigate whether fusing their representations yields additional benefit. Figure 3 compares unimodal and multimodal performance when CONCH is used as the image backbone and PCA/SCVI as the omics backbone. The complete results can be found in Table 5 in the Appendix.

Between different fusion strategies, LateMIL is more consistent across tasks than CONTACT and MCAT. Their comparison with unimodal methods is mixed: multimodal fusion outperforms the best unimodal baseline on some tasks, but there are also cases where it shows no appreciable gain or even underperforms a unimodal model. For instance, LateMIL achieve the highest AUC on BC-LOH (under CONCH+PCA) and LateMIL remains competitive on the NSCLC-Biopsy Site task. On BC-Subtype task, methods reach broadly comparable accuracy and AUC, with fusion offering marginal gains over the strongest unimodal baseline. In contrast, on the NSCLC-TMB task, GeneMLP (PCA) attains the highest ACC, with MCAT and LateMIL falling notably below it. This suggests that when a single modality carries the dominant predictive signal, fusion can weaken rather than strengthen the representation. The same pattern holds under UNI+PCA, where the relative ranking of unimodal and multimodal methods again varies across tasks. These results indicate that multimodal fusion is not universally beneficial and its utility depends on the relative informativeness of each modality for the target task.

5. Experiments on Uncertainty Quantification

Point predictions are insufficient for high-stakes oncology decision support systems, as models can still make high-confidence errors. Uncertainty quantification (UQ) addresses this by providing calibrated confidence estimates alongside each prediction (Begoli et al., 2019a). We analyse uncertainty using split conformal prediction (CP) (Angelopoulos & Bates, 2021), a model-agnostic framework

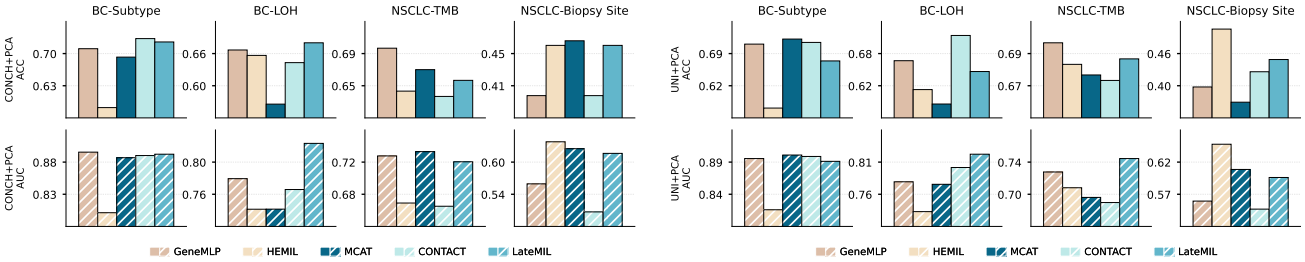

Figure 3. Performance Comparisons of Unimodal (GeneMLP, HEMIL) and Multimodal (MCAT, CONTACT, LateMIL) Methods

Table 2. Task-level conformal performance at $\alpha = 0.10$, averaged over all models.

Dataset	Task	Coverage	Cov. Gap	Avg. Set	Singleton	Accuracy	Rescue
BC	LOH	0.9021	+0.0021	1.986	0.236	0.596	75.7%
	Subtype	0.9017	+0.0017	1.995	0.310	0.641	72.6%
NSCLC	Biopsy Site	0.9309	+0.0309	2.553	0.014	0.456	87.3%
	Tumor Site	0.9154	+0.0154	2.507	0.033	0.444	84.8%

Table 3. Conformal performance by model at $\alpha = 0.10$. Each cell: Coverage / Avg. Set Size.

Dataset	Task	CONTACT	HEMIL	LateMIL	MCAT	GeneMLP
BC	LOH	0.8955 / 1.971	0.9010 / 2.016	0.9081 / 1.967	0.9014 / 1.981	0.9060 / 2.079
	Subtype	0.8981 / 1.870	0.9120 / 2.338	0.8969 / 1.994	0.9059 / 1.927	0.9025 / 2.111
NSCLC	Biopsy Site	0.9159 / 2.522	0.9402 / 2.549	0.9420 / 2.616	0.9325 / 2.493	0.9156 / 2.611
	Tumor Site	0.9149 / 2.509	0.9258 / 2.481	0.9098 / 2.514	0.9216 / 2.487	0.8986 / 2.584

that wraps any trained classifier and produces a coverage-guaranteed prediction set $\mathcal{C}(x)$ rather than a single top-1 class, applied to four tasks spanning diverse label structures: LOH, Biopsy Site, Subtype, and Tumor Site. We report results at $\alpha = 0.10$ (90% coverage target); additional α values are in Appendix C.1. We evaluate conformal prediction using three metrics: empirical coverage, average set size, and singleton rate. Empirical coverage measures how often the true label lies in the prediction set, while average set size and singleton rate quantify how informative and specific those sets are. Further details are provided in Appendix B.4.

As shown in Table 2, all four multiclass tasks achieve mean coverage at or above the nominal 0.90 target, and none exhibits a negative aggregate coverage gap. The two NSCLC endpoints deviate most strongly from the nominal target: NSCLC Biopsy Site and NSCLC Tumour Site have mean coverages of 0.931 and 0.915, respectively. This over-coverage is directly linked to their near-zero singleton rates; models are rarely confident enough to commit to a single class, so prediction sets routinely span two or three classes, driving coverage well above the nominal level at the cost of efficiency. This behaviour is consistent with the intrinsic difficulty of discriminating between lung subregions in routine clinical samples, which often yields broader, more overlapping probability distributions and therefore larger, more conservative conformal sets. The breast cancer (BC) tasks are substantially tighter. BC-LOH and BC-Subtype lie within 0.003 of the target coverage at the aggregate level, with average set sizes of approximately 2.0 out of three and four classes, respectively, indicating a more favourable reliability-efficiency trade-off. This tighter behaviour aligns with the higher predictive accuracy on BC tasks and suggests that the available image and omics features provide a stronger signal for distinguishing LOH status and molecular subtypes than for separating lung subregions, enabling models to concentrate probability mass more sharply and yield more efficient prediction sets. The rescue rate (how often the true label is recovered when the top-1 prediction is wrong) reinforces these results. NSCLC tasks benefit most from larger sets (87.3% for Biopsy Site, 84.8% for Tumour Site), while BC tasks still recover the true label in over 72% of failures, confirming that conformal prediction provides a meaningful safety net across all tasks.

Table 3 disaggregates these results by model family. For any given task, the results are more stable, the spread in average prediction-set size across the five families is less than 0.5, indicating that efficiency is largely architecture-agnostic. On the BC tasks, where coverage is closest to the nominal target, CONTACT produces the most compact sets but falls marginally below 0.90 on both LOH and Subtype, with LateMIL also under-covering slightly on Subtype. HEMIL is the most conservative family: it attains the highest coverage on BC Subtype (0.912) but at the cost of the largest average set size (2.34), consistent with a more dispersed softmax distribution than other models. For the NSCLC tasks, all families substantially exceed the target coverage, in line with the task-level analysis. The main source of variation is LateMIL, which reaches 0.942 on NSCLC Biopsy Site but drops to 0.910 on NSCLC Tumor Site, while GeneMLP is the only family to approach the lower bound on NSCLC Tumor Site. Taken together, these family-level patterns suggest that conformal behaviour is largely robust to the choice of backbone: architectural differences induce only modest changes in coverage and set size, which are small compared with the much larger task-specific effects.

6. Conclusion

We evaluated foundation model representations for multimodal cancer analysis on two real-world IH cohorts across three parts: unimodal probing, image-omics representations fusion, and methods trustworthiness. The results show

that image FMs transfer reasonably to unseen data, while the classic PCA baseline still matches or even beats omics FMs in performance. Fusion helps when both modalities contribute to the prediction, but it can underperform when the signal from one modality dominates. CP provides an architecture-agnostic uncertainty layer in which task difficulty, rather than model choice, governs prediction-set size and coverage, with prediction sets consistently recovering the true label when the top-1 prediction fails. Exploring new representation alignment methods and unified multimodal FMs to improve the performance of multimodal methods would be helpful future directions.

Acknowledgments

Portions of this research were conducted with the advanced computing resources provided by the CSCoE Converge platform. We also thank the CSCoE Scientific Computing staff for technical support. We thank Cyrus Manuel and Evan Liu for generating and providing access to the image embeddings.

Impact Statement

By coupling predictive evaluation with conformal uncertainty quantification, the study supports a shift from single point predictions to prediction sets with marginal coverage guarantees that preserve the true diagnosis in the majority of failure cases, reducing the risk of silently discarding the correct label and supporting safer human–AI collaboration in clinician decision support.

References

Abdi, H. and Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL https://arxiv.org/abs/2107.07511.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961, 2019.

Bagegni, N. A., Davis, A. A., Clifton, K. K., and Ademuyiwa, F. O. Targeted treatment for high-risk early-stage triple-negative breast cancer: Spotlight on pembrolizumab. *Breast Cancer: Targets and Therapy*, Volume 14:113–123, April 2022. ISSN 1179-1314. doi: 10.2147/bctt.s293597. URL http://dx.doi.org/10.2147/BCTT.S293597.

Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, January 2019a. ISSN 2522-5839. doi: 10.1038/s42256-018-0004-1. URL http://dx.doi.org/10.1038/s42256-018-0004-1.

Begoli, E., Bhattacharya, T., and Kusnezov, D. The need for uncertainty quantification in machine-learning for clinical applications. *Nature Machine Intelligence*, 1(1):20–23, 2019b.

Bendidi, I., Whitfield, S., Kenyon-Dean, K., Yedder, H. B., Mesbahi, Y. E., Noutahi, E., and Denton, A. K. Benchmarking transcriptomics foundation models for perturbation analysis: one pca still rules them all. *arXiv preprint arXiv:2410.13956*, 2024.

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.

Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., et al. Dna methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018.

Caris Life Sciences. Mi cancer seek. carislifesciences.com, a. URL https://www.carislifesciences.com/physicians/physician-tests/mi-cancer-seek/. Accessed May 31, 2026.

Caris Life Sciences. Mi profile. carislifesciences.com, b. URL https://www.carislifesciences.com/physicians/physician-tests/mi-profile/. Accessed May 31, 2026.

Caris Life Sciences. Mi tumor seek hybrid. carislifesciences.com, c. URL https://www.carislifesciences.com/physicians/physician-tests/mi-tumor-seek-hybrid/. Accessed May 31, 2026.

Cheerla, A. and Gevaert, O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019.

Chen, H., Venkatesh, M. S., Gómez Ortega, J., Mahesh, S. V., Nandi, T. N., Madduri, R. K., Pelka, K., and Theodoris, C. V. Scaling and quantization of large-scale foundation model enables resource-efficient predictions in network biology. *Nature Computational Science*, pp. 1–14, 2026.

Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F., Rodig, S. J., Lindeman, N. I., and Mahmood, F. 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, 2020.

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M., and Mahmood, F. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021.

Chen, R. J., Lu, M. Y., Williamson, D. F., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer cell*, 40 (8):865–878, 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*, 30(3): 850–862, 2024.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Dawood, T., Chen, C., Sidhu, B. S., Ruijsink, B., Gould, J., Porter, B., Elliott, M. K., Mehta, V., Rinaldi, C. A., Puyol-Antón, E., Razavi, R., and King, A. P. Uncertainty aware training to improve deep learning model calibration for classification of cardiac mr images. *Medical Image Analysis*, 88:102861, August 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102861. URL http://dx.doi.org/10.1016/j.media.2023.102861.

Deng, Z., Tang, C., Huang, Z., Lin, J., Chen, Y., Ning, J., Ma, C., Liu, J., Li, W., Zhu, Y., et al. Project imaging-x: A survey of 1000+ open-access medical imaging datasets for foundation model development. *arXiv preprint arXiv:2603.27460*, 2026.

Dey, S., Banerji, C. R. S., Basuchowdhuri, P., Saha, S. K., Parashar, D., and Chakraborti, T. Generating crossmodal gene expression from cancer histopathology improves multimodal ai predictions. (arXiv:2502.00568), February 2025. doi: 10.48550/arXiv.2502.00568. URL http://arxiv.org/abs/2502.00568. arXiv:2502.00568 [cs].

Ding, T., Wagner, S. J., Song, A. H., Chen, R. J., Lu, M. Y., Zhang, A., Vaidya, A. J., Jaume, G., Shaban, M., Kim, A., et al. A multimodal whole-slide foundation model for pathology. *Nature medicine*, pp. 1–13, 2025.

Duvieusart, B., Krones, F., Parsons, G., Tarassenko, L., Papież, B. W., and Mahdi, A. Multimodal cardiomegaly classification with image-derived digital biomarkers. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 13–27. Springer, 2022.

Flatiron Health. Database characterization guide. Flatiron.com, March 2025. URL https://flatiron.com/database-characterization. Published March 18, 2025. Accessed May 31, 2026.

Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., and Abbeel, P. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017a. URL http://arxiv.org/abs/1706.04599.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017b.

Hemker, K., Simidjievski, N., and Jamnik, M. Healnet: multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37: 64479–64498, 2024.

Hermessi, H., Mourali, O., and Zagrouba, E. Multimodal medical image fusion review: Theoretical background and recent advances. *Signal Processing*, 183:108036, 2021.

Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3 (1):136, 2020a.

Huang, S.-C., Pareek, A., Zamanian, R., Banerjee, I., and Lungren, M. P. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):22147, 2020b.

Jaume, G., Vaidya, A., Chen, R. J., Williamson, D. F., Liang, P. P., and Mahmood, F. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11579–11590, 2024.

Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., de Ridder, J., van Herpen, C., Lolkema, M. P., et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature communications*, 11(1):728, 2020.

Karasikov, M., van Doorn, J., Känzig, N., Erdal Cesur, M., Horlings, H. M., Berke, R., Tang, F., and Otálora, S. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 573–583. Springer, 2025.

Kedzierska, K. Z., Crawford, L., Amini, A. P., and Lu, A. X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Krones, F., Marikkar, U., Parsons, G., Szmul, A., and Mahdi, A. Review of multimodal machine learning approaches in healthcare. *Information Fusion*, 114:102690, 2025.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1): 17816, 2017.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L. P., Gerber, G., et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024a.

Lu, M. Y., Chen, B., Williamson, D. F., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., Patel, A., et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024b.

Ma, X., Long, L., Moon, S., Adamson, B. J., and Baxi, S. S. Comparison of population characteristics in real-world clinical oncology databases in the us: Flatiron health, seer, and npcr. *MedRxiv*, pp. 2020–03, 2020.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788): 89–94, 2020.

Naidoo, R., Fourkioti, O., Vries, M. D., and Bakal, C. Survivmil: A multimodal, multiple instance learning pipeline for survival outcome of neuroblastoma patients. In Ciompi, F., Khalili, N., Studer, L., Poceviciute, M., Khan, A., Veta, M., Jiao, Y., Haj-Hosseini, N., Chen, H., Raza, S., Minhas, FayyazZlobec, I., Burlutskiy, N., Vilaplana, V., Brattoli, B., Muller, H., Atzori, M., Raza, S., and Minhas, F. (eds.), *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 254 of *Proceedings of Machine Learning Research*, pp. 131–141. PMLR, 06 Oct 2024. URL https://proceedings.mlr.press/v254/naidoo24a.html.

Peeters, D., Alves, N., Venkadesh, K. V., Dinnessen, R., Saghir, Z., Scholten, E. T., Schaefer-Prokop, C., Vliegenthart, R., Prokop, M., and Jacobs, C. Enhancing a deep learning model for pulmonary nodule malignancy risk estimation in chest ct with uncertainty estimation. *European Radiology*, 34(10):6639–6651, 2024. ISSN 1432-1084. doi: 10.1007/s00330-024-10714-7.

Placido, D., Yuan, B., Hjaltelin, J. X., Zheng, C., Haue, A. D., Chmura, P. J., Yuan, C., Kim, J., Umeton, R., Antell, G., et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*, 29(5):1113–1122, 2023.

Rosen, Y., Roohani, Y., Agarwal, A., Samotorčan, L., Consortium, T. S., Quake, S. R., and Leskovec, J. Universal cell embeddings: A foundation model for cell biology. *BioRxiv*, pp. 2023–11, 2023.

Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. *NeuroImage*, 195:11–22, 2019.

Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Sellergren, A., Gao, C., Mahvar, F., Kohlberger, T., Jamil, F., Traverse, M., Tono, A., Sadjad, B., Yang, L., Lau, C., et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y. K., Kunz, J. D., Retamero, J. A., Oakley, G., Klimstra, D., Kanan, C., et al. Prism: A multi-modal generative foundation model for slide-level histopathology. *arXiv preprint arXiv:2405.10254*, 2024.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., et al. Virchow: A million-slide digital pathology foundation model. *arXiv preprint arXiv:2309.07778*, 2023.

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.

Xiang, J., Wang, X., Zhang, X., Xi, Y., Eweje, F., Chen, Y., Li, Y., Bergstrom, C., Gopaulchan, M., Kim, T., et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Zahari, R., Cox, J., and Obara, B. Quantifying the uncertainty in 3d ct lung cancer images classification. In *2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS)*, pp. 1–7, 2023. doi: 10.1109/ICPRS58416.2023.10179053.

Zhang, C., Yang, Z., He, X., and Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

Zhang, Q., Gossai, A., Monroe, S., Nussbaum, N. C., and Parrinello, C. M. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the united states. *Health services research*, 56(6):1281–1287, 2021.

A. Related Work

A.1. Medical Foundation Models

Foundation models (FMs) for computational pathology can be broadly categorized into tile-level and whole-slide-level approaches. Tile-level FMs learn patch representations from histology crops, with representative examples including contrastive learning based CTransPath (Wang et al., 2022), self-supervised model UNI (Chen et al., 2024), CONtrastive learning from Captions for Histopathology (CONCH (Lu et al., 2024a)), and Virchow (Vorontsov et al., 2023). Slide-level FMs aggregate features into holistic slide embeddings and the examples include TITAN (Ding et al., 2025) and PRISM (Shaikovski et al., 2024). For transcriptional FMs, UCE (Rosen et al., 2023) learns universal cell embeddings across species, scGPT (Cui et al., 2024) pretrains on over 33 million cells and Geneformer (Theodoris et al., 2023; Chen et al., 2026) encodes rank-value gene tokens for network-level prediction.

Beyond using FMs as feature extractors, a growing line of work explores FMs as interactive generators. PathChat (Lu et al., 2024b) pairs a pathology vision encoder with a large language model to enable conversational diagnosis, while MedGemma (Sellergren et al., 2025; 2026) extends biomedical image-text reasoning to broader medical domains. Despite these advances, recent studies (Kedzierska et al., 2025; Karasikov et al., 2025) have also shown concerns that biomedical FMs can suffer from hallucinations, misdiagnosis, modality misalignment, and in some settings, representations that underperform simpler or classical baselines. This motivates us a closer examination of how FM representations behave on downstream classification tasks.

A.2. Multimodal Learning for Downstream Tasks

Multi-modal learning methods can be categorized into early fusion, intermediate fusion, late fusion, and hybrid ways by the stage of modalities combinations (Krones et al., 2025; Huang et al., 2020a; Hermessi et al., 2021). Early fusion merges raw or minimally processed inputs via feature concatenation or stacking before they enter the main model. (Duvieusart et al., 2022) use XGBoost on image-derived biomarkers together with ICU tabular data (vital sign values, laboratory values and metadata) for cardiomegaly classification. Another practice is to use an unsupervised encoder to compress data from different modalities (e.g., clinical records and gene expression) into a single feature vector before feeding it to a survival prediction network (Cheerla & Gevaert, 2019). Intermediate fusion first extracts modality-specific representations through separate encoders, then combines them in a shared feature space. Common examples include Pathomic Fusion (Chen et al., 2020) and MCAT (Chen et al., 2021). The main idea of these methods involves first learning unimodal feature representations through individual encoders, which are then fused via mechanisms to train the final model (Krones et al., 2025). Late fusion operates at the decision level, combining independent per-modality predictions through voting, averaging, or learned meta-classifiers (Huang et al., 2020b). Hybrid methods mix multiple stages, with examples including PORPOISE (Chen et al., 2022), HEALNet (Hemker et al., 2024) and SurvPath (Jaume et al., 2024).

A.3. Uncertainty Quantification in Medical AI

Deploying machine learning in high-stakes medical settings requires not only high predictive accuracy, but also reliable indicators of when a prediction can be trusted (Begoli et al., 2019b). Early work applied approximate Bayesian methods, notably Monte Carlo dropout and deep ensem-

bles, to clinical imaging, demonstrating that uncertainty estimates can flag unreliable predictions across a range of tasks, including diabetic retinopathy screening (Leibig et al., 2017), brain lesion segmentation (Roy et al., 2019), and lung cancer and skin lesion analysis (Zahari et al., 2023; Peeters et al., 2024). Calibration has emerged as a complementary concern; models trained with cross-entropy are systematically overconfident, making post-hoc recalibration a standard practice (Guo et al., 2017b).

Conformal prediction (CP) (Angelopoulos & Bates, 2021) provides finite-sample, distribution-free coverage guarantees without assumptions on the underlying model or data-generating process, making it well suited to heterogeneous clinical datasets. Recent work has extended CP from classification to risk-bounded interval estimation; Angelopoulos et al. (2022) shows how multimodal survival models can be wrapped to yield prediction intervals with a bounded false-coverage rate, and Dey et al. (2025) applies CP to PathGen-based multimodal predictors to obtain calibrated grade sets and survival-risk intervals in computational pathology and transcriptomics.

B. Implementation Details

This appendix details the implementation of our benchmark of data preparation and exploratory analysis of IH data (§B.1), representation extraction and learning methods used in experiments (§B.2), evaluation metrics for utility performance (§B.3), and conformal prediction for uncertainty quantification (§B.4).

B.1. In-house (IH) Datasets

This subsection describes how IH-BC and IH-NSCLC cohorts are prepared for our experiments. We first specify notations and preprocessing steps, and then report exploratory statistics that summarize the distribution of the IH dataset. The data that support the findings of this study were originated by and are the property of Flatiron Health, Inc. and Caris Life Sciences. Requests for data sharing by license or by permission for the specific purpose of replicating results in this manuscript can be submitted to PublicationsDataAccess@flatiron.com and cmdb-caris@flatiron.com.

B.1.1. Data Preprocessing

Both collected IH-BC and IH-NSCLC datasets follow the format $\mathcal{D} = \{(\mathbf{x}_{i,\text{img}}, \mathbf{x}_{i,\text{omics}}, y_i)\}_{i=1}^{N}$, describing multimodal information from N cases, each indexed by a unique `case_id` $i \in \{1, \ldots, N\}$. Here $y_i \in \mathcal{Y}$ is the label for the downstream classification task, where $\mathcal{Y}$ denotes the task-specific label space (enumerated later in Table 4). Each case i is linked to corresponding raw image data $\mathbf{x}_{i,\text{img}}$ and raw omics data $\mathbf{x}_{i,\text{omics}}$. This study selects H&E whole slide images as image data and TPM RNA gene expression as omics data.

The dataset is split into train, calibration, validation, and test sets with an approximate ratio of 7:3:1:1. Specifically, IH-BC includes 3,747 unique case IDs, with 2,147 training, 1,000 calibration, 300 validation, and 300 test samples. IH-NSCLC includes 3,887 unique case IDs, with 2,287 training, 1000 calibration, 300 validation, and 300 test samples. To keep the sampled subset consistent with the original cohorts, we apply a stratified split to preserve data distributions.

Table 4 enumerated all eight downstream classification tasks considered in this study and their label spaces $\mathcal{Y}$ and class cardinalities $|\mathcal{Y}|$. For each downstream task, we recompute the class label distribution on the sampled subset and exclude any class whose samples account for less than 8% of the total, as such classes contain too few samples for the model to learn from effectively.

B.2. Models Implementation

This subsection details the models behind our benchmark. We first describe the frozen image and omics foundation models used to produce case-level representations (§B.2.1), then specify the unimodal and multimodal learning methods trained on top of these representations (§B.2.2), and finally list the computing resources used for all runs (§B.2.3).

B.2.1. Representation Generation Models

The experiments use frozen representations before downstream learning. For each case i, the raw image $\mathbf{x}_{i,\text{img}}$ is encoded into an image representation $\mathbf{z}_{i,\text{img}} \in \mathbb{R}^{B_i \times d_{\text{img}}}$, a bag of B_i tile embeddings, where each tile embedding is a d_{img}-dimensional vector and the bag size B_i varies across cases. The raw omics data $\mathbf{x}_{i,\text{omics}}$ is encoded into a case-level omics representation $\mathbf{z}_{i,\text{omics}} \in \mathbb{R}^{d_{\text{omics}}}$.

To obtain the image representation $\mathbf{z}_{i,\text{img}}$, each H&E whole slide image is divided into tiles, and a pretrained pathology encoder then maps each retained tile into a feature vector, forming a bag of d_{img}-dimensional tile embeddings. When multiple slides are available for the same case, all matched slide bags are used during training, while one slide bag is used during validation and testing.

We extract representations from four different foundation models (CONCH, UNI, Virchow, MUSK). CONCH-v1.5 is a pathology vision language foundation model trained on histology image and text pairs. Virchow is a large pathology foundation model trained at scale on cancer histopathology. UNI2 is a general purpose pathology encoder for H&E tile representations. MUSK is a multimodal pathology foundation model for image and text representation learning.

We evaluate three omics representation generation methods

Table 4. Statistics of IH-BC and IH-NSCLC downstream classification task classes. $|\mathcal{Y}|$ denotes the cardinality of the task-specific label space $\mathcal{Y}$, i.e., the number of distinct classes for that downstream task.

Dataset	Task	$\lvert\mathcal{Y}\rvert$	Class Names
IH BC	Biopsy Site	2	Metastatic / Primary
	Subtypes	4	HR-Positive HER2-Low / HR-Positive HER2-Negative / Indeterminate / Triple Negative
	Biomarker PR	2	False / True
	Biomarker PIK3CA	2	False / True
	LOH	3	Equivocal / High / Low
IH NSCLC	Biopsy Site	3	Lower lobe, lung / Lung, NOS / Upper lobe, lung
	Tumor Site	3	Lower lobe, lung / Lung, NOS / Upper lobe, lung
	TMB	2	TMB-High / TMB-Low

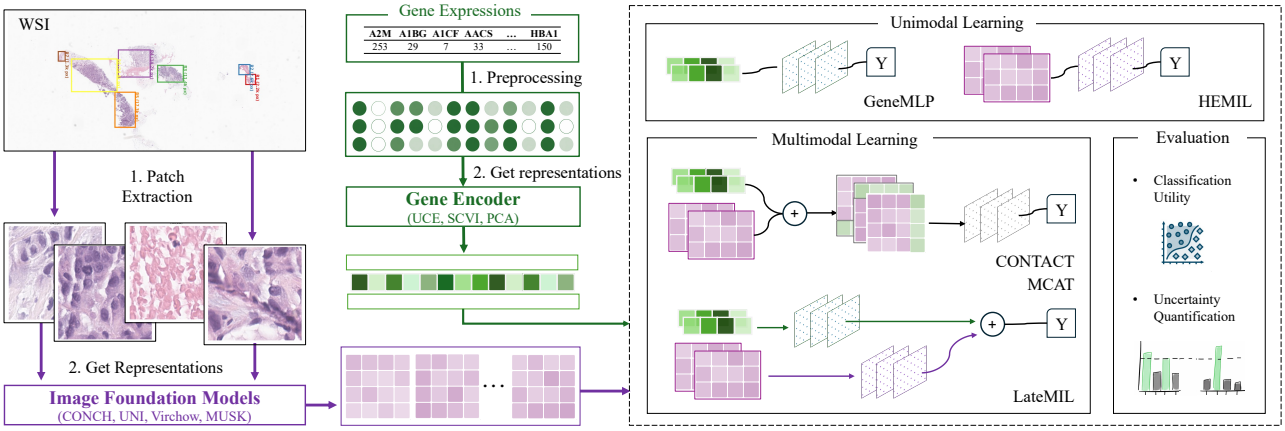

Figure 4. The Detailed Workflow

(UCE, PCA, scVI). UCE is a cell foundation model and is generates 1280-dimensional transcriptomic representations. PCA is a dimensionality reduction method that maps the gene expression matrix to 2000 principal components. scVI is a generative model based on a variational autoencoder. It is fitted with 3000 highly variable genes and provides a 1000-dimensional latent representation.

B.2.2. Representation Learning Models

Figure 4 presents a detailed workflow of this work. We consider five representation learning methods including two unimodal methods, HEMIL and GeneMLP, and three multimodal methods, CONTACT, MCAT, and LateMIL. All models are trained on the training split and final results are evaluated on the test split, with the random seed fixed to 42 throughout.

HEMIL is an H&E-based unimodal method that applies gated-attention multiple instance learning to the image representation bag $\mathbf{z}_{i,\text{img}}$. Each tile vector is projected by a linear layer with widths $\langle d_{\text{img}}, 256\rangle$ followed by GELU and dropout. The attention module has two parallel linear maps of widths $\langle 256, 256\rangle$, one with tanh and the other with sigmoid; their elementwise product is mapped to one attention logit per tile, and softmax normalizes the logits across all tiles. The bag representation is the attention-weighted sum of the projected tile vectors. The classifier head has widths $\langle 256, 128, |\mathcal{Y}|\rangle$, with LayerNorm, GELU, and dropout after the hidden layer. HEMIL is trained with weight decay=1×10^{-3} for 100 epochs and dropout 0.4. We apply early stopping with patience=20 on validation accuracy.

GeneMLP is an omics-based multilayer perceptron that receives the omics representation $\mathbf{z}_{i,\text{omics}}$ and has widths $\langle d_{\text{omics}}, 512, 256, 128, |\mathcal{Y}|\rangle$, with LayerNorm, ReLU, and dropout 0.2 after each hidden linear layer. GeneMLP is trained with weight decay=1×10^{-4} for 200 epochs. We apply early stopping with patience=20 on validation set.

CONTACT is a multimodal fusion strategy that concatenates omics information with every image tile embedding, then applies a fusion encoder with widths $\langle d_{\text{img}} + d_{\text{omics}}, 512\rangle$, with LayerNorm, GELU, and dropout after each linear layer. The fused tile bag is passed to the same MIL structure of HEMIL. CONTACT is trained with weight decay=1×10^{-2} for 100 epochs and is applied early stopping with patience=8 on validation set.

MCAT is a co-attention fusion model that exchanges information between image and omics tokens before MIL pooling. The image input projection maps each tile with widths $\langle d_{\text{img}}, 512\rangle$ using LayerNorm, GELU, and dropout; the omics input projection maps $\mathbf{z}_{i,\text{omics}}$ of width 512. Learnable modality embeddings are added to both streams, and learn-

able positional embeddings are added to the omics tokens. The encoder stacks 2 co-attention blocks: in each block, image tokens attend to omics tokens via multi-head attention, and omics tokens attend back via another multi-head attention (attention dropout 0.1); tile and omics self-attention are disabled by default. Each stream then passes through a FFN with hidden width 2048. The final image tokens are normalized and mapped with widths $\langle 512, \max(d_{\text{img}}, d_{\text{omics}})\rangle$ with LayerNorm, GELU, and dropout. The fused tile bag is classified by the same architecture of HEMIL. MCAT is trained with weight decay=1×10^{-2} for 100 epochs, dropout 0.5, omics feature dropout 0.2, label smoothing 0.05, gradient clipping at max norm 1.0. The early stopping is applied with patience=8 on validation set.

LateMIL is a late-fusion variant from (Naidoo et al., 2024) that considers both WSI logits and omics logits. The WSI branch follows a dual-stream MIL structure: the instance classifier is a linear map with widths $\langle d_{\text{img}}, |\mathcal{Y}|\rangle$ producing tile-level class logits, and the bag classifier constructs a query for each tile via a two-layer with widths $\langle d_{\text{img}}, 128, 128\rangle$, using ReLU after the first linear layer and tanh after the second. For each class, the tile with the highest instance logit is selected as the important class tile; attention scores between all tile queries and the important class tile queries are scaled and normalized over tiles, with an identity value path by default. The class-specific bag representations are passed through a convolution with kernel size d_{img} to produce WSI bag logits. The omics branch matches the architecture of GeneMLP (with LayerNorm, ReLU, and dropout 0.2). The final prediction is a fixed weighted sum of the two branch logits, with default weights 0.5. The optimizer uses separate parameter groups with WSI learning rate 2×10^{-4} and omics learning rate 2×10^{-3}, weight decay=1×10^{-4}, for 100 epochs.

B.2.3. Computing Resources

All analyses and modelling computations are performed on the data licensor's licensed compute platform. The modality learning methods are run in parallel on an HPC platform with 8 A10G GPUs with 24 GB memory each, 196 CPU cores, and 2 TB system memory.

B.3. Evaluation Metrics

Utility is evaluated on the held-out test set for each downstream classification task. We report accuracy (ACC) and area under the ROC curve (AUC) as scalar metrics, and use ROC curves to show the threshold-dependent trade-off between true positive rate (TPR) and false positive rate (FPR). Let N be the number of test cases, y_i the true label, $\hat{y}_i$ the predicted label, and p_{ic} the predicted probability assigned to class c.

Accuracy (ACC). Accuracy measures the fraction of test cases whose predicted label matches the ground-truth label, $\text{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_i = y_i\}$.

ROC curves. ROC curves evaluate probability scores over all possible classification thresholds. For a given class $c \in \mathcal{Y}$, cases with $y_i = c$ are treated as positives and all other cases are treated as negatives. At threshold τ, the true positive rate (TPR) and false positive rate (FPR) are defined as

$$\text{TPR}_c(\tau) = \frac{\sum_{i=1}^{N} \mathbf{1}\{y_i = c,\ p_{ic} \geq \tau\}}{\sum_{i=1}^{N} \mathbf{1}\{y_i = c\}},$$

$$\text{FPR}_c(\tau) = \frac{\sum_{i=1}^{N} \mathbf{1}\{y_i \neq c,\ p_{ic} \geq \tau\}}{\sum_{i=1}^{N} \mathbf{1}\{y_i \neq c\}}.$$

The ROC curve is obtained by plotting TPR against FPR as τ varies. Binary tasks use the positive-class ROC curve. Multi-class tasks use a macro one-vs-rest ROC curve, $\overline{\text{TPR}}(u) = \frac{1}{|\mathcal{Y}|}\sum_{c\in\mathcal{Y}} \text{TPR}_c(u)$, after each class-specific ROC curve is reparameterised as $\text{TPR}_c(u)$ by inverting the monotone non-increasing relation $u = \text{FPR}_c(\tau)$, eliminating τ. The diagonal line $\text{TPR} = \text{FPR}$ is included as the random-ranking reference.

Area under the ROC curve (AUC). AUC is the area under the ROC curve and summarizes how well the model ranks positive cases above negative cases using predicted probabilities rather than hard labels. For class $c \in \mathcal{Y}$, $\text{AUC}_c = \int_0^1 \text{TPR}_c(u)\, du$, where u denotes the false positive rate. For binary tasks, AUC is computed from the positive-class probability, following the positive class specified per task in Table 4. For multi-class tasks, we use a one-vs-rest macro average, $\text{AUC}_{\text{macro}} = \frac{1}{|\mathcal{Y}|}\sum_{c\in\mathcal{Y}} \text{AUC}_c$, where $\mathcal{Y}$ is the task label space. This gives equal weight to each class and avoids letting larger classes dominate the reported AUC.

B.4. Conformal Quantification

We evaluate our classifiers in an uncertainty-aware setting using split conformal prediction at miscoverage levels $\alpha \in \{0.05, 0.10, 0.20\}$. For each α, prediction sets are formed by including all classes whose nonconformity score falls below a calibrated threshold, guaranteeing marginal coverage $\Pr(y \in \mathcal{C}(x)) \geq 1 - \alpha$.

The nonconformity score for a sample x and class c is defined as

$$s(x, c) = 1 - p_\theta(c \mid x), \tag{1}$$

where $p_\theta(c \mid x)$ is the softmax probability assigned to class c. On the calibration set, each score is evaluated at the true label, i.e. $s_i = s(x, c)$. Given calibration scores $\{s_i\}_{i=1}^{n_{\text{cal}}}$,

the conformal threshold $\hat{q}$ is computed as the q-th quantile of the calibration scores, where

$$q = \frac{\lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil}{n_{\text{cal}}} \tag{2}$$

is the finite-sample corrected quantile level. A test class c is then included in the prediction set if and only if

$$s(x, c) \leq \hat{q}. \tag{3}$$

We evaluate conformal prediction results using the following metrics.

Coverage measures the fraction of test samples for which the true label falls inside the prediction set, and is the primary metric for verifying that the nominal guarantee is satisfied,

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i \in \mathcal{C}(x_i)\}. \tag{4}$$

Target coverage is the nominal coverage level imposed by the choice of miscoverage rate α, and it is used as the reference against which empirical coverage is compared $\text{Target} = 1 - \alpha$..

Coverage gap quantifies the deviation of empirical coverage from the nominal target. Positive values indicate over-coverage; negative values indicate a violation of the guarantee: $\text{CoverageGap} = \text{Coverage} - (1 - \alpha)$.

Average prediction set size measures prediction efficiency, how many classes must the model include, on average, to achieve the required coverage. Smaller values indicate sharper, more informative predictions:

$$\text{AvgSetSize} = \frac{1}{n} \sum_{i=1}^{n} |\mathcal{C}(x_i)|. \tag{5}$$

Singleton rate measures the fraction of samples for which the model is confident enough to return exactly one class. High singleton rates indicate decisive, well-concentrated predictions:

$$\text{SingletonRate} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{|\mathcal{C}(x_i)| = 1\}. \tag{6}$$

We also report the exact achievable coverage level δ_{exact}, which differs from the nominal target $\delta_{\text{target}} = 1 - \alpha$ due to the discreteness of the calibration quantile:

$$\delta_{\text{exact}} = \frac{\lceil (n_{\text{cal}} + 1)\, \delta_{\text{target}} \rceil}{n_{\text{cal}} + 1}. \tag{7}$$

The slack $\delta_{\text{exact}} - \delta_{\text{target}}$ quantifies the unavoidable over-coverage introduced by finite-sample rounding, while the empirical gap $\text{Coverage} - \delta_{\text{target}}$ measures the total deviation from the nominal target. Together, these diagnostics allow us to distinguish genuine coverage violations from artefacts of finite calibration sets.

Expected Calibration Error (ECE) (Guo et al., 2017a) quantifies how well the model's predicted probabilities agree with empirical accuracies. The ECE is then defined as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|\beta_m|}{n} \left| \text{acc}(\beta_m) - \text{conf}(\beta_m) \right|, \tag{8}$$

where $|\beta_m|$ is the number of samples in bin β_m and n is the total number of samples. This measures the average, sample-weighted discrepancy between predicted confidence and realised accuracy across bins.

Brier score (Dawood et al., 2023) provides a complementary measure of probabilistic accuracy and calibration. It is defined as

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2, \tag{9}$$

where N is the number of samples, p_i is the predicted probability vector in the multi-class case, for sample i, and y_i is the corresponding one-hot encoded ground-truth label. Lower Brier scores indicate that the predicted probabilities are closer to the observed outcomes, reflecting better-calibrated and more accurate probabilistic predictions.

To quantify the practical advantage of conformal prediction over a single point estimate, we computed the **rescue rate**: among all test samples with an incorrect top-1 prediction, this is the fraction for which the true class still appears in the conformal prediction set ($\alpha = 0.10$). A high rescue rate indicates that, even when the model is wrong, the correct answer is preserved among the plausible candidates, reducing the risk of silently discarding the true diagnosis in a clinical workflow.

C. Appendix: Full Results

Table 5 shows detailed results for all multimodal techniques with different image and omics representations backbones.

Overall, models on BC tasks show better predictive performance than NSCLC tasks. This difference may reflect multiple factors such as task definition, label noise, and demographic composition. A subgroup analysis would be needed to isolate demographic effects: the BC models are effectively built on a more demographically homogeneous population than the NSCLC models. Due to the nature of the disease, BC dataset we extracted are all female patients, so the model is trained and evaluated on a single-gender cohort. In contrast, NSCLC dataset contains a more balanced distribution of male and female patients, but demographic information is excluded from modelling. Similar clinical presentations can correspond to different classification labels across genders, making the prediction task inherently more challenging. Incorporating demographic information into the modelling process and evaluating variations in predictive performance across subgroups would be beneficial for building more robust predictive models.

Across different downstream tasks, we observe a pattern that task labels anchored in tissue context tend to favor image representations, whereas task labels anchored in molecular state tend to favor omics representations. For instance, the image-based HEMIL probe overall outperforms the omics-based GeneMLP probe on the Biopsy Site identification tasks for both NSCLC and BC datasets. This may be because biopsy site is primarily determined by surrounding tissue context and tumor morphology, both of which favor image-based probing. By contrast, biomarker-status tasks, such as BC-PR, target molecular states whose primary substrate is transcriptional, and therefore favor omics-leaning probing. As a result, GeneMLP outperforms HEMIL on BC-PR status prediction.

Table 5. Multimodal Fusion Performance on IH-BC and IH-NSCLC Tasks. Each cell reports ACC / AUC.

Dataset	Img-modality	CONCH	UNI	MUSK	CONCH	UNI	MUSK	CONCH	UNI	MUSK
	Omics-modality	PCA	PCA	PCA	UCE	UCE	UCE	SCVI	SCVI	SCVI
BC LOH	GeneMLP	0.6667 / 0.7794	0.6667 / 0.7794	0.6667 / 0.7794	0.5253 / 0.6419	0.5253 / 0.6419	0.5253 / 0.6419	0.4882 / 0.7287	0.4882 / 0.7287	0.4882 / 0.7287
	HEMIL	0.6566 / 0.7414	0.6128 / 0.7330	0.6431 / 0.7420	0.6566 / 0.7414	0.6128 / 0.7330	0.6431 / 0.7420	0.6566 / 0.7414	0.6128 / 0.7330	0.6431 / 0.7420
	MCAT	0.5657 / 0.7415	0.5859 / 0.7755	0.6027 / 0.7579	0.6465 / 0.7325	0.6195 / 0.7786	0.5589 / 0.7272	0.6364 / 0.7746	0.6027 / 0.7725	0.6094 / 0.7673
	CONTACT	0.6431 / 0.7659	0.7138 / 0.8016	0.6431 / 0.7312	0.5926 / 0.7287	0.6128 / 0.7203	0.5758 / 0.7077	0.5825 / 0.7472	0.6263 / 0.7649	0.6263 / 0.7568
	LateMIL	0.6801 / 0.8233	0.6465 / 0.8223	0.6835 / 0.7907	0.6094 / 0.7356	0.6835 / 0.7699	0.6263 / 0.7185	0.6532 / 0.7773	0.6633 / 0.7835	0.6330 / 0.7614
BC Biopsy Site	GeneMLP	0.7797 / 0.8467	0.7797 / 0.8467	0.7797 / 0.8467	0.6136 / 0.7000	0.6136 / 0.7000	0.6136 / 0.7000	0.7119 / 0.7590	0.7119 / 0.7590	0.7119 / 0.7590
	HEMIL	0.8780 / 0.9111	0.8814 / 0.9343	0.8508 / 0.9191	0.8780 / 0.9111	0.8814 / 0.9343	0.8508 / 0.9191	0.8780 / 0.9111	0.8814 / 0.9343	0.8508 / 0.9191
	MCAT	0.8169 / 0.9015	0.8508 / 0.9030	0.8034 / 0.8754	0.8610 / 0.9242	0.8814 / 0.9252	0.8407 / 0.9114	0.8508 / 0.9205	0.8712 / 0.9119	0.8237 / 0.8995
	CONTACT	0.7525 / 0.8463	0.7661 / 0.8412	0.7424 / 0.8224	0.8407 / 0.9194	0.8610 / 0.9245	0.8542 / 0.9039	0.7898 / 0.8902	0.8441 / 0.9181	0.8373 / 0.9054
	LateMIL	0.8373 / 0.9028	0.8102 / 0.9089	0.8441 / 0.9037	0.8542 / 0.9195	0.8983 / 0.9424	0.8373 / 0.9249	0.8407 / 0.9218	0.8983 / 0.9432	0.8407 / 0.9321
BC PR	GeneMLP	0.7700 / 0.8159	0.7700 / 0.8159	0.7700 / 0.8159	0.6200 / 0.6297	0.6200 / 0.6297	0.6200 / 0.6297	0.7333 / 0.7633	0.7333 / 0.7633	0.7333 / 0.7633
	HEMIL	0.6933 / 0.7288	0.7133 / 0.7139	0.6900 / 0.7453	0.6933 / 0.7288	0.7133 / 0.7139	0.6900 / 0.7453	0.6933 / 0.7288	0.7133 / 0.7139	0.6900 / 0.7453
	MCAT	0.7000 / 0.7341	0.6700 / 0.7364	0.6667 / 0.7266	0.7267 / 0.7292	0.6967 / 0.7106	0.7100 / 0.7286	0.7133 / 0.7513	0.7067 / 0.7288	0.7467 / 0.7763
	CONTACT	0.7167 / 0.7627	0.6767 / 0.7331	0.7100 / 0.7472	0.7233 / 0.7154	0.7000 / 0.7198	0.7333 / 0.7165	0.6933 / 0.7444	0.7133 / 0.7527	0.7133 / 0.7548
	LateMIL	0.7367 / 0.7890	0.7100 / 0.7744	0.7200 / 0.7666	0.7200 / 0.7252	0.6933 / 0.7005	0.6967 / 0.7258	0.7167 / 0.7611	0.7300 / 0.7483	0.7200 / 0.7633
BC PIK3CA	GeneMLP	0.7933 / 0.7921	0.7933 / 0.7921	0.7933 / 0.7921	0.6967 / 0.6097	0.6967 / 0.6097	0.6967 / 0.6097	0.7033 / 0.7718	0.7033 / 0.7718	0.7033 / 0.7718
	HEMIL	0.7500 / 0.7355	0.7433 / 0.7299	0.7533 / 0.6969	0.7500 / 0.7355	0.7433 / 0.7299	0.7533 / 0.6969	0.7500 / 0.7355	0.7433 / 0.7299	0.7533 / 0.6969
	MCAT	0.7100 / 0.7600	0.7600 / 0.7598	0.6700 / 0.7478	0.7367 / 0.7291	0.7000 / 0.7191	0.7533 / 0.7168	0.7400 / 0.7531	0.7033 / 0.7343	0.7100 / 0.7866
	CONTACT	0.7300 / 0.7179	0.7333 / 0.7615	0.7533 / 0.7328	0.7467 / 0.7053	0.7333 / 0.6577	0.7500 / 0.6834	0.7467 / 0.7592	0.7167 / 0.7724	0.7300 / 0.7524
	LateMIL	0.8100 / 0.8161	0.7733 / 0.8047	0.7533 / 0.7978	0.7000 / 0.7166	0.7433 / 0.7175	0.7433 / 0.6928	0.7400 / 0.7815	0.7400 / 0.7737	0.7367 / 0.7791
BC Subtype	GeneMLP	0.7106 / 0.8955	0.7106 / 0.8955	0.7106 / 0.8955	0.5311 / 0.7558	0.5311 / 0.7558	0.5311 / 0.7558	0.6520 / 0.8649	0.6520 / 0.8649	0.6520 / 0.8649
	HEMIL	0.5824 / 0.8013	0.5714 / 0.8159	0.5788 / 0.7986	0.5824 / 0.8013	0.5714 / 0.8159	0.5788 / 0.7986	0.5824 / 0.8013	0.5714 / 0.8159	0.5788 / 0.7986
	MCAT	0.6923 / 0.8868	0.7216 / 0.9010	0.6740 / 0.8867	0.5751 / 0.8159	0.5824 / 0.8243	0.5897 / 0.8096	0.6410 / 0.8755	0.6557 / 0.8555	0.6557 / 0.8755
	CONTACT	0.7326 / 0.8902	0.7143 / 0.8988	0.6813 / 0.8821	0.5641 / 0.8151	0.5641 / 0.8062	0.5641 / 0.8040	0.6630 / 0.8777	0.6630 / 0.8808	0.6777 / 0.8823
	LateMIL	0.7253 / 0.8923	0.6740 / 0.8912	0.6740 / 0.8786	0.5678 / 0.8015	0.6337 / 0.8418	0.5311 / 0.7680	0.6337 / 0.8597	0.6813 / 0.8759	0.6300 / 0.8458
NSCLC Biopsy Site	GeneMLP	0.3977 / 0.5594	0.3977 / 0.5594	0.3977 / 0.5594	0.4375 / 0.5053	0.4375 / 0.5053	0.4375 / 0.5053	0.3295 / 0.5205	0.3295 / 0.5205	0.3295 / 0.5205
	HEMIL	0.4602 / 0.6379	0.5057 / 0.6478	0.4830 / 0.6475	0.4602 / 0.6379	0.5057 / 0.6478	0.4830 / 0.6475	0.4602 / 0.6379	0.5057 / 0.6478	0.4830 / 0.6475
	MCAT	0.4659 / 0.6250	0.3693 / 0.6086	0.4318 / 0.6136	0.4773 / 0.6334	0.4545 / 0.6378	0.4716 / 0.6362	0.4830 / 0.6007	0.4318 / 0.5915	0.4602 / 0.6146
	CONTACT	0.3977 / 0.5071	0.4261 / 0.5469	0.3693 / 0.5116	0.5114 / 0.6537	0.4773 / 0.6305	0.4716 / 0.6039	0.4602 / 0.6320	0.4261 / 0.5886	0.3920 / 0.6123
	LateMIL	0.4602 / 0.6163	0.4489 / 0.5961	0.4318 / 0.5494	0.4602 / 0.6373	0.4716 / 0.6208	0.4432 / 0.5825	0.4773 / 0.6151	0.4830 / 0.6410	0.3580 / 0.5730
NSCLC TMB	GeneMLP	0.6967 / 0.7277	0.6967 / 0.7277	0.6967 / 0.7277	0.6433 / 0.6364	0.6433 / 0.6364	0.6433 / 0.6364	0.5967 / 0.6542	0.5967 / 0.6542	0.5967 / 0.6542
	HEMIL	0.6433 / 0.6691	0.6833 / 0.7081	0.6633 / 0.6210	0.6433 / 0.6691	0.6833 / 0.7081	0.6633 / 0.6210	0.6433 / 0.6691	0.6833 / 0.7081	0.6633 / 0.6210
	MCAT	0.6700 / 0.7330	0.6767 / 0.6961	0.6567 / 0.6723	0.6667 / 0.6775	0.6200 / 0.6711	0.6767 / 0.6588	0.6433 / 0.6604	0.6267 / 0.6709	0.6500 / 0.6802
	CONTACT	0.6367 / 0.6651	0.6733 / 0.6897	0.6733 / 0.6666	0.6600 / 0.6703	0.6667 / 0.6593	0.6600 / 0.5893	0.6567 / 0.6884	0.6633 / 0.6711	0.6700 / 0.6874
	LateMIL	0.6567 / 0.7204	0.6867 / 0.7444	0.6567 / 0.6991	0.5633 / 0.6733	0.6933 / 0.7190	0.6567 / 0.6780	0.6233 / 0.6559	0.6800 / 0.6933	0.6400 / 0.6568
NSCLC Tumor Site	GeneMLP	0.4346 / 0.5915	0.4346 / 0.5915	0.4346 / 0.5915	0.4664 / 0.6100	0.4664 / 0.6100	0.4664 / 0.6100	0.4629 / 0.6014	0.4629 / 0.6014	0.4629 / 0.6014
	HEMIL	0.4982 / 0.6283	0.4912 / 0.6520	0.4841 / 0.6177	0.4982 / 0.6283	0.4912 / 0.6520	0.4841 / 0.6177	0.4982 / 0.6283	0.4912 / 0.6520	0.4841 / 0.6177
	MCAT	0.4841 / 0.6274	0.4770 / 0.6075	0.4523 / 0.6206	0.4982 / 0.6631	0.4523 / 0.6585	0.4452 / 0.6318	0.4488 / 0.6372	0.4134 / 0.5967	0.4452 / 0.6145
	CONTACT	0.4099 / 0.5705	0.4205 / 0.5428	0.4452 / 0.5796	0.4982 / 0.6485	0.4558 / 0.6334	0.4912 / 0.6289	0.4735 / 0.6403	0.4382 / 0.6330	0.4735 / 0.6314
	LateMIL	0.4700 / 0.6027	0.4912 / 0.6318	0.4700 / 0.6169	0.4629 / 0.6342	0.4876 / 0.6346	0.4841 / 0.6054	0.4629 / 0.6301	0.5159 / 0.6598	0.4452 / 0.6315

Figure 5 reports macro-averaged ROC curves of the two unimodal baselines (GeneMLP for omics, HEMIL for WSI) across four representative tasks under different encoder backbones. Overall, the patterns across different tasks are consistent with

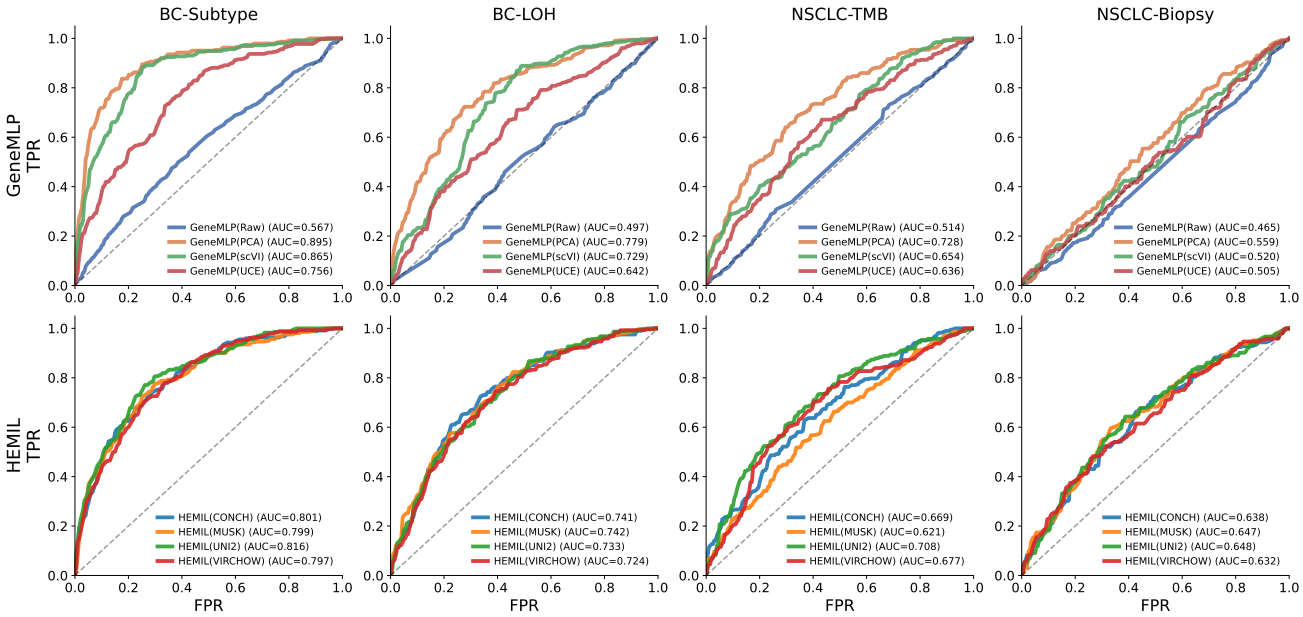

Figure 5. ROC curves of unimodal baselines across different tasks (BC-Subtype, BC-LOH, NSCLC-TMB, NSCLC-Biopsy) Here 'RAW' refers to directly feeding the full set of gene TPM values as omics features.

what we observed in the main text. The various image-based foundation models show stable predictive performance. For omics representations, all three encoded versions outperform using raw gene expression directly. That said, the omics foundation model UCE generally underperforms PCA and scVI. This is likely because UCE is pre-trained and generates representations in a training-free manner, whereas PCA and scVI are further fitted on the training set before producing representations. Fine-tuning omics foundation models on domain-specific data in future work can help them generate more informative embeddings.

C.1. Conformal Prediction

Figure 6 extends the single-level analysis to the full range of target coverages. The same task-level structure observed at $\alpha = 0.10$ is evident across all three panels, the BC tasks track the nominal targets closely, whereas the NSCLC tasks consistently overshoot, because their harder multi-site discrimination task leads to more variable, higher non-conformity scores and thus larger, more conservative prediction sets.

At $\alpha = 0.05$ (target 0.95), BC configurations cluster tightly between 0.95 and 0.97. The only clear under-coverage is GeneMLP on NSCLC Tumour Site (coverage 0.9471, gap -0.003), while LateMIL on NSCLC Biopsy Site shows the strongest over-coverage (0.9776, gap $+0.028$; Table 6). At $\alpha = 0.10$ (target 0.90), BC families again form a narrow band around the target, whereas NSCLC families overshoot by roughly 2–4 percentage points; GeneMLP on NSCLC Tumour Site (0.8986, gap -0.001) is the only near-miss. At $\alpha = 0.20$ (target 0.80), NSCLC over-coverage remains high, while CONTACT and MCAT on BC LOH dip slightly below target (gaps -0.010 and -0.012), making this level the most prone to under-coverage on BC tasks (Table 6). The relative ordering of model families within each task remains stable and the spread of coverages does not grow systematically with α, indicating that the qualitative conclusions from the $\alpha = 0.10$ setting generalise across the coverage sweep.

Table 6 also reports probabilistic calibration via the expected calibration error (ECE) (Guo et al., 2017a) and the multi-class Brier score (Dawood et al., 2023), both of which are independent of α and therefore reflect intrinsic model properties rather than the choice of coverage level. CONTACT and MCAT are consistently well calibrated, maintaining ECE values in the range 0.04–0.07 across all four tasks, whereas LateMIL and GeneMLP are notably less reliable, with ECE reaching 0.144 and 0.160 on the BC tasks and peaking at 0.204 for GeneMLP on NSCLC Tumour Site. The Brier score follows the same path. NSCLC tasks have substantially higher values (0.62–0.77) than BC tasks (0.46–0.59), consistent with their increased difficulty, and within each endpoint, the ordering of families matches the ECE ranking, with CONTACT and MCAT achieving the lowest Brier scores and GeneMLP the highest.

These probabilistic calibration mirrors the conformal results. Models with broad, poorly concentrated softmax distributions,

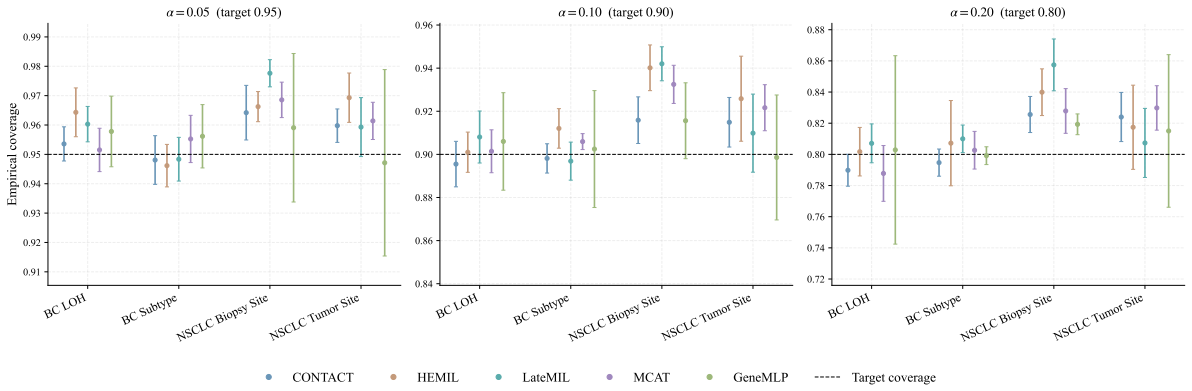

Figure 6. Empirical coverage across models and target coverage levels $\alpha \in \{0.05, 0.10, 0.20\}$. Each point is the mean coverage averaged over encoder configurations for a given model and task.

Table 6. Conformal prediction metrics per task and model family across coverage levels ($\alpha \in \{0.05, 0.10, 0.20\}$). Each α-block reports: empirical coverage (Cov.), coverage gap $\Delta = \text{cov} - (1 - \alpha)$ (Gap), average prediction-set size (Set), and singleton rate (Sing.). Fixed columns (right): expected calibration error (ECE), multi-class Brier score, and top-1 accuracy (Acc.), which are independent of α. Results averaged over encoder configurations.

Task	Model	$\alpha = 0.05$				$\alpha = 0.10$				$\alpha = 0.20$				ECE	Brier	Acc.
		Cov.	Gap	Set	Sing.	Cov.	Gap	Set	Sing.	Cov.	Gap	Set	Sing.			
BC – LOH	CONTACT	0.9536	+0.0036	2.35	0.097	0.8955	-0.0045	1.97	0.249	0.7898	-0.0102	1.53	0.500	0.0575	0.510	0.609
	HEMIL	0.9643	+0.0143	2.48	0.081	0.9010	+0.0010	2.02	0.251	0.8017	+0.0017	1.55	0.499	0.0661	0.510	0.613
	LateMIL	0.9603	+0.0103	2.25	0.138	0.9081	+0.0080	1.97	0.271	0.8071	+0.0071	1.57	0.504	0.1444	0.545	0.599
	MCAT	0.9515	+0.0015	2.22	0.093	0.9014	+0.0014	1.98	0.194	0.7878	-0.0122	1.56	0.454	0.0620	0.515	0.579
	GeneMLP	0.9578	+0.0078	2.37	0.049	0.9060	+0.0060	2.08	0.158	0.8029	+0.0029	1.68	0.397	0.1596	0.593	0.573
BC – Subtype	CONTACT	0.9481	-0.0019	2.36	0.178	0.8981	-0.0019	1.87	0.350	0.7947	-0.0053	1.38	0.638	0.0531	0.463	0.654
	HEMIL	0.9462	-0.0038	2.73	0.068	0.9120	+0.0120	2.34	0.163	0.8072	+0.0072	1.70	0.424	0.1242	0.559	0.593
	LateMIL	0.9484	-0.0016	2.55	0.187	0.8969	-0.0031	1.99	0.347	0.8100	+0.0100	1.51	0.576	0.1250	0.507	0.646
	MCAT	0.9553	+0.0052	2.57	0.067	0.9059	+0.0059	1.93	0.315	0.8027	+0.0027	1.40	0.621	0.0549	0.466	0.652
	GeneMLP	0.9562	+0.0062	2.60	0.108	0.9025	+0.0025	2.11	0.247	0.7992	-0.0008	1.54	0.556	0.1240	0.523	0.621
NSCLC – Biopsy Site	CONTACT	0.9642	+0.0142	2.71	0.002	0.9159	+0.0159	2.52	0.007	0.8256	+0.0256	2.17	0.044	0.0508	0.630	0.453
	HEMIL	0.9663	+0.0163	2.72	0.003	0.9402	+0.0402	2.55	0.015	0.8399	+0.0399	2.10	0.132	0.1195	0.651	0.482
	LateMIL	0.9776	+0.0276	2.81	0.006	0.9420	+0.0420	2.62	0.025	0.8574	+0.0574	2.23	0.100	0.1482	0.671	0.455
	MCAT	0.9686	+0.0186	2.69	0.000	0.9325	+0.0325	2.49	0.007	0.8279	+0.0279	2.09	0.054	0.0436	0.623	0.463
	GeneMLP	0.9591	+0.0091	2.78	0.008	0.9156	+0.0156	2.61	0.017	0.8193	+0.0193	2.26	0.066	0.1349	0.695	0.406
NSCLC – Tumor Site	CONTACT	0.9598	+0.0098	2.72	0.005	0.9149	+0.0149	2.51	0.031	0.8240	+0.0240	2.16	0.105	0.0598	0.636	0.433
	HEMIL	0.9693	+0.0193	2.75	0.010	0.9258	+0.0258	2.48	0.047	0.8174	+0.0174	2.03	0.147	0.0754	0.623	0.474
	LateMIL	0.9593	+0.0093	2.77	0.014	0.9098	+0.0098	2.51	0.046	0.8073	+0.0073	2.10	0.156	0.1231	0.674	0.444
	MCAT	0.9614	+0.0114	2.68	0.004	0.9216	+0.0216	2.49	0.015	0.8298	+0.0298	2.13	0.079	0.0679	0.633	0.453
	GeneMLP	0.9471	-0.0029	2.77	0.009	0.8986	-0.0014	2.58	0.023	0.8150	+0.0150	2.26	0.082	0.2037	0.773	0.397

as indicated by high ECE and Brier, require larger $\hat{q}$ thresholds to include the true class, increasing prediction-set sizes and producing over-coverage. This is most evident for GeneMLP on NSCLC Tumour Site, where high ECE (0.204) and Brier (0.773) coincide with the largest prediction sets in that task (average size ≈ 2.58–2.77 across α levels; Table 6) and with the only near under coverage at $\alpha = 0.05$ (coverage 0.947, gap -0.003). Conversely, MCAT on NSCLC Biopsy Site combines the lowest ECE in that task (0.044) with the smallest prediction sets, illustrating how better probabilistic calibration translates directly into tighter and more informative prediction sets.

Figure 7 illustrates the top-1 error, which counts how often the single most likely class is wrong, and the conformal miss rate, which counts how often the true label falls outside the prediction set $\mathcal{C}(x)$. The difference between these two reflects the uncertainty absorbed by the prediction set, where the classifier misclassifies the top label but the conformal wrapper still includes the true class. Across the BC tasks (top row), top-1 error ranges from approximately 35% to 43% depending on the model family, yet the conformal miss rate remains close to the nominal 10% target for all five families. The hatched bars are short and nearly identical, indicating that the choice of architecture has virtually no impact on coverage reliability at this level. The NSCLC tasks (bottom row) show a different pattern. Top-1 error increases to 52–60%, reflecting the harder multi-site discrimination, while conformal miss rates fall below the target to about 7–8%, consistent with the systematic over-coverage reported in Table 2. The gap between solid and hatched bars is therefore largest on the NSCLC tasks, the

conformal wrapper absorbs substantially more uncertainty, covering many samples that the point predictor misclassifies. This gap is also highly consistent across model families, suggesting that the score distributions of all architectures are broad enough that the true label typically remains inside $\mathcal{C}(x)$ regardless of backbone choice.

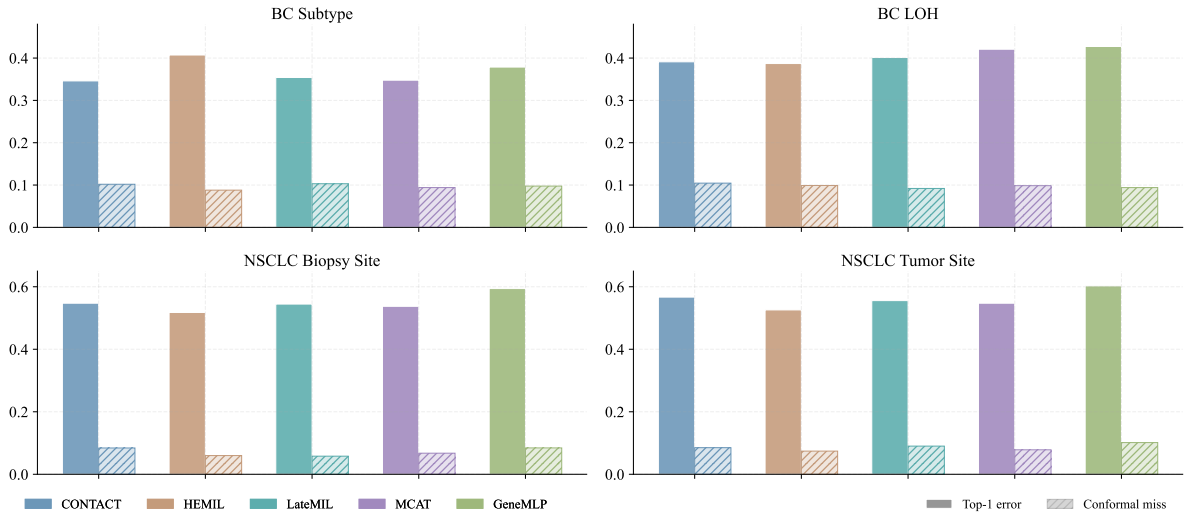

Figure 7. Error decomposition by task and model family. Each panel shows one of the four multiclass tasks; within each panel, bars are grouped by model family (colour). Solid bars represent the top-1 classification error ($1 -$ accuracy); hatched bars represent the conformal miss rate ($1 -$ coverage) at $\alpha = 0.10$. The gap between the two bars quantifies the uncertainty absorbed by the prediction set, samples misclassified by the point predictor that are nonetheless covered by $\mathcal{C}(x)$ because the true class retains a softmax score above $1 - \hat{q}$.

From the rescue rates in Table 6, we observe that across all four multiclass tasks and all model configurations, 79.7% of misclassified samples were *rescued*, meaning that the conformal set still contained the correct label even when the single-best prediction was wrong. The rescue rate varied by clinical task: 72.6% for BC Subtype, 75.7% for BC LOH, 87.3% for NSCLC Biopsy Site, and 84.8% for NSCLC Tumour Site. These numbers indicate that, in the large majority of failure cases, a hard classifier would commit to an incorrect label, whereas conformal prediction would still flag the true diagnosis as a plausible outcome and defer the final decision to a human reviewer.

Two patient-level examples from the BC Subtype task illustrate the clinical impact of this behaviour. In the first example, a HEMIL model assigned a softmax probability of 0.368 to HR+/HER2-Low and 0.364 to the true class HR+/HER2-Negative, a very small margin of $\Delta p = 0.004$. A standard classifier would pick HR+/HER2-Low and be wrong.[1] A false HR+/HER2-Low label would therefore wrongly suggest that the patient is eligible for T-DXd and could send them down an inappropriate treatment pathway. By contrast, the conformal prediction set at $\hat{q} = 0.860$ was $\mathcal{C}(x) = \{$HR+/HER2-Low, HR+/HER2-Neg, Indeterminate$\}$, a three-class output that clearly shows HER2 status is uncertain and that additional IHC or FISH testing is needed before deciding on treatment.

In the second example, the same model assigned a probability of 0.336 to Indeterminate and 0.331 to the true class, Triple Negative, again a tiny margin ($\Delta p = 0.004$). Triple Negative Breast Cancer (TNBC) is defined by the absence of some receptors and is managed very differently from an Indeterminate case. Patients with TNBC may receive intensive systemic therapies such as chemotherapy. An Indeterminate prediction would instead trigger further biomarker workup and delay the start of TNBC-directed therapy, which is problematic given the aggressive course of the disease (Bagegni et al., 2022). The conformal prediction set $\mathcal{C}(x) = \{$HR+/HER2 Low, Indeterminate, Triple Neg.$\}$ keeps Triple Negative as an active diagnostic option alongside Indeterminate, encouraging clinicians to pursue both possibilities in parallel rather than committing to a single ambiguous label. Taken together, these examples show that conformal prediction sets do more than improve statistical coverage: they translate model uncertainty into concrete clinical prompts for individual patients.

[1] See, for example, the FDA approval summary for fam-trastuzumab deruxtecan-nxki (Enhertu) in HER2-low breast cancer: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-fam-trastuzumab-deruxtecan-nxki-her2-low-breast-cancer.