SYNPRED: A Synergistic Approach to Multimodal Learning for Clinical Prediction

Ivana Janíčková[1,2,5], Yen Y. Tan[3], Thomas H. Helbich[6], Philipp Seeböck[1,2], Julieta Di Marco[1,2], Zsuzsanna Bago-Horvath[4], Ulrike Heber[4], Thomas Spiegel[6], and Georg Langs[1,2,5]

[1] Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria
[2] Comprehensive Center for Artificial Intelligence in Medicine, Medical University of Vienna, Austria
[3] Department of Obstetrics and Gynecology, Medical University of Vienna, Austria
[4] Department of Pathology, Medical University of Vienna, Austria
[5] Christian Doppler Laboratory for Machine Learning Driven Precision Imaging, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria
[6] Division of General and Pediactric Radiology, Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria
ivana.janickova@meduniwien.ac.at, georg.langs@meduniwien.ac.at
https://www.cir.meduniwien.ac.at

Abstract. Multimodal medical data captures complementary aspects of disease. It provides a comprehensive basis for diagnosis and the identification of biological associations linked to progression or treatment response. Established contrastive learning approaches exploit shared information across modalities, but fall short in exploiting complementary, modality-specific signals relevant for a task. Here, we propose a synergistic multimodal variational autoencoder (VAE) framework that separates task-relevant unimodal and multimodal information from instance-specific variation task-irrelevant, allowing complementary signals across modalities to be leveraged for prediction. The method yields benefits even if modalities are missing at test time. We evaluate the method on a treatment response prediction task using medical imaging and RNA data. Empirical comparison with alignment-based and multimodal VAE state-of-the-art approaches demonstrates improved multimodal prediction performance, and consistent benefits for prediction from a single modality after multimodal training. The full code can be found here: https://github.com/cirmuw/SYNPRED.

Keywords: Multimodal Data · Variational Inference · Representation Learning.

1 Introduction

Multimodal data can improve predictive performance and clinical decision support in medical machine learning [9]. Importantly, it often combines modalities

describing complementary aspects of disease, sharing only limited semantic overlap. For instance, radiological imaging may reflect tumour structure, whereas biopsy-derived molecular data characterises underlying biological processes such as methylation or gene expression [7, 12]. Multimodal data integration approaches based on alignment (contrastive or attention-based) are fundamentally limited, as they primarily capture shared information and fail to explicitly model complementary, modality-specific signals that nevertheless might be informative for the prediction task [5].

Here we propose a multimodal generative framework that learns structured latent representations for clinical prediction. Instead of representing only characteristics shared across modalities, our model transparently disentangles (1) shared and (2) complementary (modality-specific) task-relevant features, and (3) residual features independent of the prediction target. We show that explicitly modeling shared and complementary task-relevant features enables the model to capture synergistic cross-modal information —information that cannot be recovered from any single modality alone [5, 23] - during multimodal training, improving downstream multimodal and single-modality prediction.

Related Work. Multimodal representation learning includes variational approaches for shared representation learning and handling of missing modalities [14, 23, 24], as well as alignment-based contrastive and attention-based methods [21, 11, 4, 17]. In medical applications, most methods focus on data with strong semantic overlap and rely on contrastive alignment or attention mechanisms [27, 29, 10, 3]. As noted by [5], such approaches primarily emphasize shared information and may under represent complementary or modality-specific task-relevant signals, particularly when semantic correspondence between modalities is limited. Classical radiomics and radiogenomics methods address multimodal data with lower semantic overlap via feature correlation [7, 12]. Deep learning approaches typically rely on late fusion through feature concatenation [26, 8]. Only a small number of recent contributions propose more structured solutions for such data, including alignment-based [25] and attention-based approaches [1, 15]. Despite their differing mechanisms, these methods remain limited in their ability to handle missing modalities and to explicitly model complementary, modality-specific task information. Recent analyses highlight the importance of such complementary information, that multimodal pretraining can uncover *synergistic* cross-modal patterns beyond shared information. While synergy arises from multimodal interactions, learning these patterns during training can enhance unimodal representations and improve single-modality performance in downstream tasks [23, 5].

Contribution. We propose SYNPRED, a synergistic multimodal VAE for representation learning in radiology-genomics settings with limited semantic overlap between modalities. Unlike alignment-based approaches that primarily emphasize shared information, our framework explicitly disentangles shared task-relevant (*fusion*), modality-specific task-relevant (*uni*), and task-irrelevant instance-specific (*inst*) latent components, enabling complementary signals to contribute to pre-

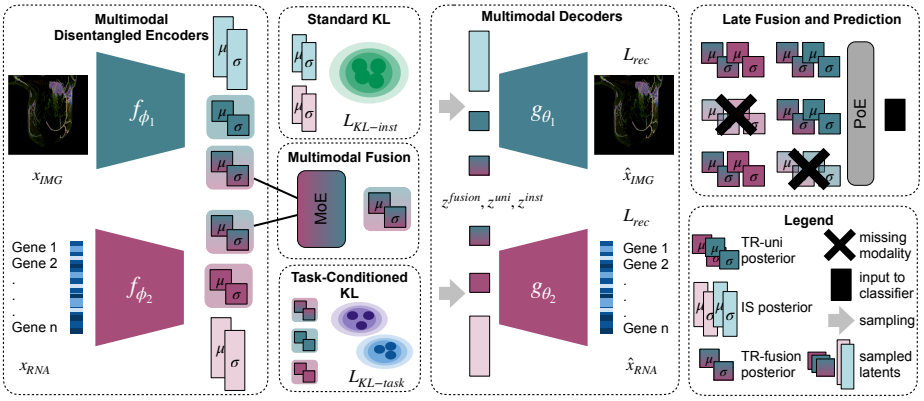


Fig. 1: **Method Overview** Each modality-specific encoder maps the input to a factorised Gaussian posterior, decomposing the latent space into shared task-relevant (TR-fusion), modality-specific task-relevant (TR-uni), and instance-specific (IS) components with posterior parameters (μ, σ) and sampled latents (z). TR-fusion posteriors of each modality are combined via a mixture-of-experts (MoE). Task-conditioned KL regularisation ($\mathcal{L}_{\text{KL-task}}$) is applied to TR-fusion and TR-uni posteriors, while a KL term ($\mathcal{L}_{\text{KL-inst}}$) regularises IS posteriors. The sampled latents z are concatenated and decoded using a reconstruction loss ($\mathcal{L}_{\text{rec}}$). For prediction, embeddings are combined using a product-of-experts (PoE).

diction. We evaluate the framework on the public BreastDCEDL dataset [6], combining paired breast MRI and gene expression data from the I-SPY2 cohort [16, 19]. Experiments show improved multimodal performance and consistent gains in single-modality prediction, particularly for imaging, providing empirical evidence that multimodal training improves representation learning through synergistic cross-modal interactions.

2 Method

We propose **SYNPRED**, a **syn**ergistic **pred**iction model for learning disentangled, task-relevant representations from medical modalities using a multimodal VAE framework. As illustrated in Fig. 1, each modality is modeled using multimodal task-relevant information (*fusion*), modality-specific task-relevant information (*uni*), and instance-specific variation (*inst*). The model is trained with reconstruction and variational regularization under standard and task-conditioned priors, enabling robust prediction under missing modalities via mixture- and product-of-expert fusion. The following problem formulation formalizes this structure.

Problem Formulation We consider a two-modality setting with inputs X_1, X_2 and a task label Y. The goal is to predict Y from multimodal observations by

exploiting shared and complementary information across modalities. Motivated by the decomposition of multimodal information into redundant and complementary components [2], we assume that each modality is generated from instance-specific variation, modality-specific task-relevant information, and shared task-relevant information. We model this structure using disentangled latent representations that preserve complementary and shared signals, enabling synergistic task-relevant information to emerge.

2.1 SYNPRED: Synergistic Multimodal VAE

Latent Decomposition and Reconstruction We instantiate the framework for images and gene expression data using one encoder–decoder pair per modality and a shared latent space. Each encoder f_{ϕ_m} maps modality input x_m to factorised Gaussian posteriors over shared task-relevant (TR-fusion), modality-specific task-relevant (TR-uni), and instance-specific (IS) latent components:

$$q_{\phi_m}(z_m^{\text{fusion}}, z_m^{\text{uni}}, z_m^{\text{inst}} \mid x_m) = \underbrace{q_{\phi_m}(z_m^{\text{fusion}} \mid x_m)}_{\text{TR-fusion}} \underbrace{q_{\phi_m}(z_m^{\text{uni}} \mid x_m)}_{\text{TR-uni}} \underbrace{q_{\phi_m}(z_m^{\text{inst}} \mid x_m)}_{\text{IS}}. \tag{1}$$

The total dimensionality is $d = 2d_t + d_i$, where d_t and d_i denote the dimensionalities of task-relevant and instance-specific components, respectively. For each latent component z_m^p, the encoder outputs the mean and variance of the approximate posterior

$$q_{\phi_m}(z_m^p \mid x_m) = \mathcal{N}\big(\mu_m^p(x_m), \operatorname{diag}(\sigma_m^p(x_m)^2)\big), \qquad p \in \{\text{fusion}, \text{uni}, \text{inst}\}. \tag{2}$$

The decoder g_{θ_m} reconstructs $\hat{x}_m = g_{\theta_m}(z_m^{\text{fusion}}, z_m^{\text{uni}}, z_m^{\text{inst}})$ for $m \in \{1, 2\}$ from the concatenated sampled latents. Reconstructions are optimised using an ℓ_1 loss, while instance-specific latents (IS) are regularised toward a standard Normal prior using the Kullback–Leibler (KL) divergence:

$$\mathcal{L}_{\text{rec}}^{(m)} = \mathbb{E}[\|\hat{x}_m - x_m\|_1], \qquad \mathcal{L}_{\text{KL-inst}}^{(m)} = \mathrm{KL}\big(q_{\phi_m}(z_m^{\text{inst}} \mid x_m) \,\|\, \mathcal{N}(0, I)\big). \tag{3}$$

Multimodal Fusion Fusion task-relevant posteriors (TR-fusion) of each modality $q_{\phi_m}(z_m^{\text{fusion}} \mid x_m)$ are combined using mixture-of-experts (MoE) [23] over the set of available modalities $\mathcal{M}$. The parameters of multimodal fused posterior $q(z^{\text{fusion}} \mid x_1, x_2)$ are computed as weighted averages of the modality-specific posteriors, with learnable mixture weights w_m for each modality:

$$\mu^{\text{fusion}} = \sum_{m \in \mathcal{M}} w_m \mu_m^{\text{fusion}}, \qquad \sigma^{\text{fusion}} = \sum_{m \in \mathcal{M}} w_m \sigma_m^{\text{fusion}}. \tag{4}$$

Task Conditioning The task-relevant posteriors (TR-fusion, TR-uni) $q_{\phi_m}(z_m^p \mid x_m)$, $p \in \{\text{fusion}, \text{uni}\}$, are shaped using an observed target label y. Inspired by [13], we condition the generative model on y by defining task-conditional

Gaussian priors: $p(z^p \mid y) = \mathcal{N}\big(\mu_y^p,\ \mathrm{diag}(\sigma_y^{p2})\big)$, which are used to construct the task-conditioned KL loss $\mathcal{L}_{\text{KL-task}}$. Here, separate parameters $\{\mu_y^p, \sigma_y^p\}$ are learned for each latent type p and target label y, encouraging the z^{fusion} latents to capture shared task-related information and the z^{uni} latents to capture modality-specific complementary task-related information. Task conditioning is enforced via KL divergence terms between the approximate posteriors and the corresponding task-conditional priors:

$$\begin{aligned}\mathcal{L}_{\text{KL-task}} = \mathrm{KL}\big(q(z^{\text{fusion}} \mid x_1, x_2) \parallel \mathcal{N}\big(\mu_y^{\text{fusion}},\ \mathrm{diag}(\sigma_y^{\text{fusion}2})\big)\big) \\ + \mathrm{KL}\big(q_{\phi_m}(z_m^{\text{uni}} \mid x_m) \parallel \mathcal{N}\big(\mu_y^{\text{uni}},\ \mathrm{diag}(\sigma_y^{\text{uni}2})\big)\big)\,.\end{aligned} \tag{5}$$

Overall Loss The model is trained by minimizing a weighted sum of reconstruction and the two KL-based terms with β is a fixed hyperparameter controlling the strength of latent regularisation:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \beta\,(\mathcal{L}_{\text{KL-inst}} + \mathcal{L}_{\text{KL-task}})\,, \tag{6}$$

Prediction For downstream prediction, embeddings are formed by joining posterior parameters of the shared task-relevant (TR-fusion) and modality-specific task-relevant (TR-uni) latents, $(\mu_m, \log\sigma_m^2) = ([\mu_m^{\text{fusion}}; \mu_m^{\text{uni}}], [\log\sigma_m^{2,\text{fusion}}; \log\sigma_m^{2,\text{uni}}])$. Modality-specific Gaussian experts are combined via a product-of-experts (PoE), and the fused mean is used as the classifier input. Representations are evaluated via linear probing and fine-tuning, trained on multimodal embeddings and tested on unimodal or multimodal inputs without retraining.

3 Experiments and Results

Data We use the multi-center public BreastDCEDL dataset [6], which contains pretreatment breast cancer (BC) patients with a binary pathological complete response (pCR) label (1452 patients). The dataset aggregates dynamic contrast-enhanced (DCE) MRI from the I-SPY1 [18], I-SPY2 [16, 19], and Duke [22] cohorts, and we adopt the official train/validation/test split. For multimodal experiments, the I-SPY2 cohort (982 patients) additionally provides paired pretreatment gene-expression profiles (GEO: GSE194040) [20, 28], containing measurements for over 19000 genes. From the DCE MRI, we compute a 2D maximum-intensity projection of early and late subtraction images. Images are intensity-normalised to $[0, 1]$. RNA features are mean-imputed, log-transformed, and z-score normalised. During training, we apply simple noise-based augmentation, using Gaussian and Gibbs noise for images and Gaussian noise for RNA inputs.

Implementation Details The model consists of two modality-specific VAE branches and a fusion module. The image branch consists of convolutional, and the RNA branch of fully-connected layers with hidden size 256. Latent variables are partitioned into *fusion, uni* and *inst* components, with total dimensionality

Table 1: **Comparison with SOTA:** Mean AUROC and 95% confidence interval results for image/RNA-only, and multimodal evaluation.† and ‡ indicate complete modality requirement during pretraining and fine-tuning respectively. All baselines were finetuned, for SYNPRED (lin.prob.) is linear probing; (ft) is finetuning.

Method		IMG-only AUROC	RNA-only AUROC	Multimodal AUROC
Single-Modal Approaches				
IMG VAE		0.549 [0.508,0.590]	—	—
RNA VAE		—	0.656 [0.614,0.702]	—
Multimodal Approaches				
concat	‡	—	—	0.644 [0.630,0.657]
ContIG [25]	†	0.560 [0.553,0.566]	0.702 [0.682,0.715]	0.696 [0.681,0.709]
MMVM VAE [24]	†	0.512 [0.456,0.566]	0.634 [0.604,0.659]	0.628 [0.594,0.661]
DMVAE [14]		0.542 [0.534,0.550]	0.668 [0.655,0.683]	0.670 [0.653,0.685]
SYNPRED (lin.prob.)		0.638 [0.601,0.673]	0.706 [0.666,0.741]	**0.725 [0.682,0.757]**
SYNPRED (ft)		**0.693 [0.620,0.740]**	**0.717 [0.698,0.734]**	0.707 [0.695,0.719]

$d = 256$ ($d_c = 16$ per task-related component). Multimodal fusion is performed via MoE over modality-specific fusion posteriors, with learnable softmax weights. The model is trained using separate learning rates for the image and RNA branches (10^{-5} and 10^{-6}, respectively) and a fixed regularisation weight $\beta = 10^{-4}$ in Eq. 6. Models are trained for 300 epochs (batch size 16) across five seeds, with selection based on validation AUROC. Ablations alter latent disentanglement and associated losses. Performance is evaluated via linear probing and MLP fine-tuning, with missing RNA handled through MoE and PoE fusion.

3.1 Results

Comparison with state-of-the-art approaches SYNPRED outperforms state-of-the-art (SOTA) methods in multimodal prediction of treatment response (pCR) for BC patients. Even if only a single modality is available during test time, it substantially outperforms alternative approaches. This is particularly pronounced for the prediction from imaging data where simple uni-modal approach yields AUROC of 0.549 while SYNPRED achieves and AUROC of 0.693. This suggests that RNA data can help the model to better identify predictive signatures in the imaging data. We compare against several pretraining approaches fine-tuned with an MLP classifier (Table 1), including unimodal VAEs and a simple concatenation-based multimodal baseline. For alignment-based pretraining, we use ContIG [25], a contrastive framework that enforces cross-modal representation alignment. As multimodal VAE baselines, we include MMVM VAE [24], which uses a data-dependent prior for soft multimodal information sharing, and DMVAE [14], which disentangles multimodal - neither method explicitly disentangles task-relevant from task-irrelevant information. Following prior work

Table 2: **Ablation study of latent components.** Mean AUROC (95% CI) for image-only, RNA-only, and multimodal prediction using linear probing, comparing a non-disentangled baseline and models with ablated task-relevant (*uni*, *fusion*) and instance-specific (*inst*) latent components.

Ablation			IMG-only	RNA-only	Multimodal
z^{uni}	z^{fusion}	z^{inst}	AUROC	AUROC	AUROC
✗	✗	✗	0.524 [0.488,0.559]	0.643 [0.606,0.680]	0.646 [0.611,0.682]
✓	✓	✗	0.555 [0.505,0.606]	0.674 [0.626,0.723]	0.680 [0.632,0.730]
✓	✗	✓	0.603 [0.508,0.671]	**0.706 [0.669,0.748]**	0.719 [0.693,0.760]
✗	✓	✓	0.524 [0.468,0.573]	0.647 [0.604,0.681]	0.658 [0.611,0.670]
✓	✓	✓	**0.638 [0.601,0.673]**	**0.706 [0.666,0.741]**	**0.725 [0.682,0.757]**

[24], multimodal performance for the SOTA methods is reported as the average of single-modality results; multimodal evaluation uses paired samples, while unimodal results use all available data.

Ablation Study Ablation results (Table 2) highlight the importance of disentangling task-specific (*fusion, uni*) and instance-specific (*inst*) latent. Removing disentanglement substantially degrades multimodal performance, reducing AUROC from 0.725 to 0.646, while omitting either unimodal or fused task-relevant components also leads to lower performance (0.658–0.680), indicating that both modality-specific and multimodal information are important for prediction.

Analysis of Explainability The proposed method is able to improve localization of clinically relevant regions that are informative for the prediciton task. We assess model's focus on tumour region by comparing it to saliency-based explanations and assessing the overlap with ground-truth segmentations of lesions (Fig. 2a). The proposed method achieves the highest overlap compared to SOTA approaches, reaching a DICE score of 0.167. In representation-based evaluation, the full disentangled model consistently outperforms not only the non-disentangled variant but also all ablated configurations, indicating that jointly modeling modality-specific and multimodal task-relevant information yields the most spatially coherent explainability maps.

4 Discussion

Disentangling complementary and shared task-relevant information, rather than enforcing cross-modal alignment, is important for multimodal learning in settings with limited semantic overlap, such as radiology-genomics applications combining RNA and imaging data. This is particularly relevant when modalities have asymmetric predictive strength, with imaging acting as the weaker modality. Existing multimodal generative approaches for medical data [25, 1, 15] primarily emphasize

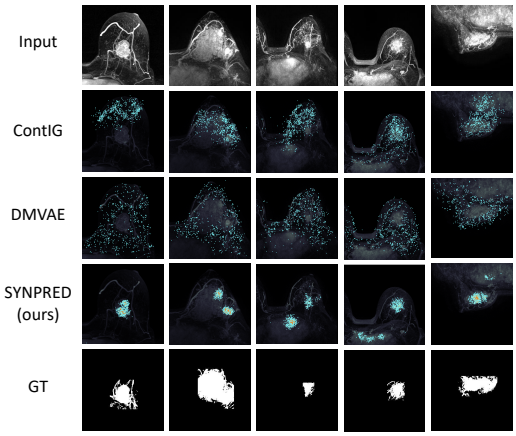
shared information, often underutilizing modality-specific yet task-relevant signals. In contrast, our framework explicitly disentangles shared, modality-specific, and instance-specific representations to better capture complementary multimodal information.

Empirical results suggest that disentanglement and explicit modeling of modality-specific and multimodal components contribute to robust prediction and clinically meaningful localization. The proposed method achieves stronger performance than alignment-based and multimodal VAE baselines in multimodal prediction (Table 1). Importantly, multimodal training also improves single-modality prediction, particularly for imaging, which has lower standalone predictive power and benefits most from complementary RNA information. These gains provide empirical evidence that synergistic cross-modal interactions during multimodal training can uncover predictive structure in the weaker imaging modality that is difficult to capture from imaging alone. Ablation studies (Table 2) further indicate that both the *uni* and *fusion* components contribute to performance, while explainability analysis shows improved overlap with ground-truth segmentations compared to alignment-based and VAE baselines. While applicable in principle, the framework was not evaluated on settings with more than two modalities or on tasks beyond classification. In addition, the current evaluation is limited to a single radiology-genomics benchmark involving paired breast MRI and RNA data, and broader validation on additional datasets is needed to assess generalizability.

Prediction-based	
Method	DICE
ImgVAE	0.043±0.023
ContIG [25]	0.080±0.003
MMVM VAE [24]	0.007±0.002
DMVAE [14]	0.016±0.004
SYNPRED (ft)	**0.167±0.002**

Representation-based			
z^{uni}	z^{fusion}	z^{inst}	DICE
✗	✗	✗	0.069±0.088
✗	✓	✓	0.121±0.141
✓	✓	✗	0.113±0.016
✓	✗	✓	0.120±0.147
✓	✓	✓	**0.133±0.128**

(a) Quantitative Results



(b) Qualitative Results

Fig. 2: **Explainability evaluation.** (*Left*) DICE overlap between saliency maps and ground-truth segmentations. (*Right*) Qualitative saliency visualizations. The proposed method achieves the highest overlap and more coherent localization of task-relevant regions.

5 Conclusion

We present a multimodal deep generative framework for clinical prediction from medical modalities with limited semantic overlap. By explicitly disentangling modality-specific and shared task-relevant representations from instance-specific variation, the proposed approach improves representation learning from complementary multimodal signals. Empirical results show consistent gains in single-modality prediction after multimodal training, particularly for imaging with limited standalone predictive power, while maintaining competitive multimodal performance. These findings provide empirical evidence that synergistic cross-modal interactions during multimodal training can uncover predictive structure in the weaker imaging modality that would otherwise remain difficult to capture.

Acknowledgments. This work was funded by the Vienna Science and Technology Fund (WWTF, PREDICTOME [10.47379/LS20065]) the European Union's Horizon Europe research and innovation programme under grant agreement No.101100633— EUCAIM, Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association, and Siemens Healthineers.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ai, Y., Liu, J., Li, Y., Wang, F., Du, X., Jain, R.K., Lin, L., Chen, Y.W.: Sama: a self-and-mutual attention network for accurate recurrence prediction of non-small cell lung cancer using genetic and ct data. IEEE Journal of Biomedical and Health Informatics **29**(5), 3220–3233 (2024)
2. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., Ay, N.: Quantifying unique information. Entropy **16**(4), 2161–2183 (2014)
3. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., 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)
4. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
5. Dufumier, B., Castillo Navarro, J., Tuia, D., Thiran, J.P.: What to align in multimodal contrastive learning? In: International Conference on Learning Representations. vol. 2025, pp. 5408–5432 (2025)
6. Fridman, N., Solway, B., Fridman, T., Barnea, I., Goldstein, A.: Breastdcedl: A standardized deep learning-ready breast dce-mri dataset of 2,070 patients. Scientific Data (2026)
7. Guan, J., Fan, M., Li, L.: Mvnmf: Multiview nonnegative matrix factorization for radio-multigenomic analysis in breast cancer prognosis. Medical Image Analysis **103**, 103566 (2025)

8. Huang, J.X., Shi, J., Ding, S.S., Zhang, H.L., Wang, X.Y., Lin, S.Y., Xu, Y.F., Wei, M.J., Liu, L.Z., Pei, X.Q.: Deep learning model based on dual-modal ultrasound and molecular data for predicting response to neoadjuvant chemotherapy in breast cancer. Academic Radiology **30**, S50–S61 (2023)
9. Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., 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 (2020)
10. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3942–3951 (2021)
11. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
12. Kawahara, D., Kishi, M., Kadooka, Y., Hirose, K., Murakami, Y.: Integrating radiomics and gene expression by mapping on the image with improved deepinsight for clear cell renal cell carcinoma. Cancer genetics **292**, 100–105 (2025)
13. Kingma, D.P., Rezende, D.J., Mohamed, S., Welling, M.: Semi-supervised learning with deep generative models. Advances in neural information processing systems **27** (2014)
14. Lee, M., Pavlovic, V.: Private-shared disentangled multimodal vae for learning of latent representations. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 1692–1700 (2021)
15. Li, H., Zhao, Y., Duan, J., Gu, J., Liu, Z., Zhang, H., Zhang, Y., Li, Z.C.: Mri and rna-seq fusion for prediction of pathological response to neoadjuvant chemotherapy in breast cancer. Displays **83**, 102698 (2024)
16. Li, W., Newitt, D., Gibbs, J., Wilmes, L., Jones, E., Arasu, V., Strand, F., Onishi, N., Nguyen, A., Kornak, J., et al.: I-spy 2 breast dynamic contrast enhanced mri trial (ispy2)(version 1). Cancer Imaging Arch **10** (2022)
17. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
18. Newitt, D., Hylton, N.: Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials. the cancer imaging archive (2016)
19. Newitt, D., Partridge, S., Zhang, Z., Gibbs, J., Chenevert, T., Rosen, M., Bolan, P., Marques, H., Romanoff, J., Cimino, L., et al.: Acrin 6698/i-spy2 breast dwi [data set](2021). DOI: https://doi. org/10.7937/tcia. kk02-6d95 (2025)
20. Petricoin, E.F., Wolf, D.M., Yau, C., Wulfkuhle, J.D., van't Veer, L., Gallagher, R.I., Esserman, L.J., Hirst, G.L., Brown-Swigart, L., Yee, D., et al.: Immune and growth factor signaling pathways are associated with pathologic complete response to an anti–type i insulin-like growth factor receptor regimen in patients with breast cancer. Clinical Cancer Research **31**(20), 4361–4371 (2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
22. Saha, A., Harowicz, M.R., Grimm, L.J., Weng, J., Cain, E.H., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. The Cancer Imaging Archive **10** (2021)

23. Shi, Y., Paige, B., Torr, P., et al.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in neural information processing systems **32** (2019)
24. Sutter, T., Meng, Y., Agostini, A., Chopard, D., Fortin, N., Vogt, J., Shahbaba, B., Mandt, S.: Unity by diversity: Improved representation learning for multimodal vaes. Advances in Neural Information Processing Systems **37**, 74262–74297 (2024)
25. Taleb, A., Kirchler, M., Monti, R., Lippert, C.: Contig: Self-supervised multimodal contrastive learning for medical imaging with genetics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20908–20921 (2022)
26. Tripathi, S., Moyer, E.J., Augustin, A.I., Zavalny, A., Dheer, S., Sukumaran, R., Schwartz, D., Gorski, B., Dako, F., Kim, E.: Radgennets: Deep learning-based radiogenomics model for gene mutation prediction in lung cancer. Informatics in medicine unlocked **33**, 101062 (2022)
27. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2022, p. 3876 (2022)
28. Wolf, D.M., Yau, C., Wulfkuhle, J., Brown-Swigart, L., Gallagher, R.I., Lee, P.R.E., Zhu, Z., Magbanua, M.J., Sayaman, R., O'Grady, N., et al.: Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies. Cancer cell **40**(6), 609–623 (2022)
29. Yan, R., Zhang, X., Jiang, Z., Wang, B., Bian, X., Ren, F., Zhou, S.K.: Pathway-aware multimodal transformer (pamt): Integrating pathological image and gene expression for interpretable cancer survival analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)