Graph Representation Learning of Longitudinal Medical Imaging Trajectories for Treatment Response Prediction

Johannes Kiechle[1,2,3,7,‡], Richard Osuala[2,4], Daniel M. Lang[1,3], Stefan M. Fischer[1,2,3,7], Ivana Janíčková[5], Karim Lekadir[4,8], Julia A. Schnabel[1,3,6,7,†], and Jan C. Peeken[2,†]

[1] Technical University of Munich, [2] TUM University Hospital Rechts der Isar, [3] Helmholtz Munich, [4] Universitat de Barcelona, [5] Medical University of Vienna, [6] King's College London, [7] Munich Center for Machine Learning, [8] Institució Catalana de Recerca i Estudis Avançats (ICREA)
‡johannes.kiechle@tum.de

Abstract. In patients with breast cancer, pathological complete response (pCR) has been established as a clinically meaningful surrogate marker for long-term outcomes. While commonly treated with neoadjuvant chemotherapy (NACT), effective treatment decision-making remains challenging, as therapeutic response can vary substantially across patients, calling for predictive models capable of accurately estimating individualized treatment response. To address this, we propose an imaging-based 3D spatio-temporal framework for treatment response prediction that integrates a state-of-the-art graph neural network with relational modeling of temporal interactions across timepoints alongside three novel and complementary treatment trajectory representation learning objectives. Experiments across a cohort of 585 patients from the public ISPY2 dataset demonstrate that our method substantially outperforms both supervised and supervised representation learning baselines across several classification metrics. Alongside establishing a breast cancer pCR prediction benchmark, we include a principled ablation of our method and further introduce and empirically assess the impact of the available number of DCE-MRI timepoints per patient trajectory and the inclusion of inter-scan time-differences. Overall, our study substantiates the utility of clinically meaningful longitudinal medical imaging modeling for predicting NACT-induced pCR. The code accompanying this work (https://github.com/compai-lab/2026-miccai-grail-kiechle) alongside a user-friendly Python library for data curation (https://pypi.org/project/ISPY2-pCR-dataset), is publicly available.

Keywords: GNN · Longitudinal · Supervised Representation Learning

† Shared senior authorship.

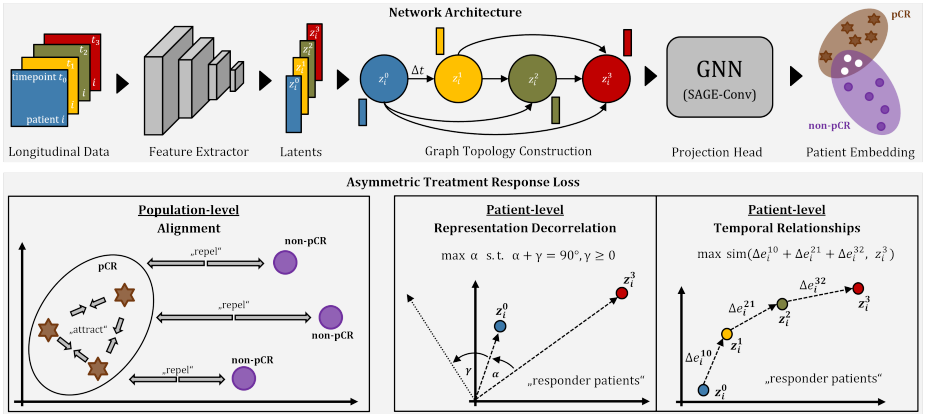

Fig. 1. Method: Longitudinal image-derived latent features are represented as a directed acyclic graph and aggregated through a graph neural network (GNN) projection head to obtain a compact patient-level embedding. Representation learning is guided by an asymmetric response-aware objective that combines (a) population-level alignment, (b) patient-level representation decorrelation, and (c) temporal consistency losses.

1 Introduction

With an estimated 2.3 million women diagnosed with breast cancer annually [1], optimizing treatment remains a clinical priority where neoadjuvant chemotherapy (NACT) crucially enables early response assessment to guide therapy adaptation and surgical planning [2]. Pathologic complete response (pCR), i.e. the absence of residual invasive disease in the breast and axillary lymph nodes at surgery following NACT [2], is commonly used to estimate NACT efficacy while being associated with improved long-term outcomes and event-free survival [3].

Reliable early prediction of pCR holds the promise of mitigating treatment-related toxicity and crucially supports informed clinical decision-making, including therapeutic escalation or de-escalation strategies. Moreover, early identification of non-responders may enable timely treatment modification and optimization of surgical intervention. Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) plays an essential role in screening and assessment of response to NACT [4]. Accordingly, DCE-MRI–based methods for pCR prediction have attracted substantial attention. Many existing studies, however, are limited by the use of single-modality pre-treatment data, absence of rigorous baseline examinations, or reliance on single train-test split, thereby constraining confidence in generalizability of reported performances [4, 2, 5–11]. While imaging data from a single time point (e.g. pre-treatment) insufficiently captures subtle therapy-induced alterations [8], the integration of longitudinal DCE-MRI acquired over the course of NACT [12, 13] allow pCR prediction models to capture dynamic tumor evolution, signaling individual treatment trajectories.

Nonetheless, only a limited number of studies have utilized longitudinal DCE-MRI data [11], and existing approaches remain constrained in their ability to learn complex temporal representations. Several methods rely on hand-crafted features [7] or simple concatenation of multiple time points [10], which may inadequately capture longitudinal tumor dynamics. For example, [14] incorporated multiple DCE-MRI time points but employed a fixed CNN pretrained on natural images for feature extraction, without adapting representations to the temporal treatment setting. Similarly, [8] applied an LSTM to model inter-timepoint dependencies, however, training was restricted to a supervised binary pCR classification objective, without dedicated temporal representation learning. Likewise, Zang et al. [15] proposed a temporal fusion module integrating attention-based and sequence modeling mechanisms to aggregate longitudinal MRI features, yet optimization was again limited to supervised pCR classification.

Recently, a purely imaging-based 2D longitudinal representation learning approach [11] demonstrated promising results but evaluated performance on a single dataset split, likely overestimating generalization performance. While pointing towards the needs of (i) deeper multi-split validation and (ii) CNN-based feature extractor and supervised representation learning baseline comparisons [11, 8, 16], these previous studies signaled potential capabilities of temporal representation learning of disease trajectories. Although these studies suggest potential, its downstream impact, specifically for breast cancer pCR prediction, including varying numbers of DCE-MRI time points remains insufficiently explored [11, 17]. We aim to bridge this gap by introducing a time-aware graph neural network with novel temporal representation learning objectives, alongside a principled pCR prediction framework and benchmark evaluation. The main contributions of this work are summarized as follows:

1. We propose a novel time-aware graph neural network that explicitly encodes structured treatment response trajectories within its graph topology.
2. We introduce three novel representation learning objectives that enforce trajectory and response-consistent latent representations.
3. We establish a comprehensive imaging-based benchmark for breast cancer pCR prediction that evaluates different methods, available imaging timepoints, and the inclusion of inter-scan temporal differences.

2 Methods and Materials

2.1 Network Architecture

Visual Feature Extraction Given a 3D input volume, representing a patient i at timepoint t, the longitudinal MRI set is described as $\{x_i^t\}_{t=1}^T$, where $x_i \in \mathbb{R}^{C\times D\times H\times W}$. Each input volume is independently processed by a ResNet [18] backbone to extract corresponding feature maps, which are subsequently passed through a global average pooling layer, to obtain a compact embedding vector representing an 3D input volume of patient i at timepoint t, denoted as $z_i^t \in \mathbb{R}^D$.

Graph Topology Construction The set of image embeddings $\{z_i^t\}_{t=1}^T$ representing a patient i across T timepoints is subsequently translated into a directed graph, where each node corresponds to one embedding $z_i^t \in \mathbb{R}^D$ and edges encode pairwise relationships. Therein, nodes are indexed according to their temporal ordering. For every pair of nodes (u, v) with $u < v$, we add a directed edge from node u to node v, with no reverse edges, resulting in a set of edges $\mathcal{E} = \{(u, v) \mid 0 \leq u < v < N\}$. This yields a directed acyclic graph, which ensures that each node can propagate information to all subsequent nodes while preserving temporal dependencies. The choice of this topology ensures pairwise node connectivity under a strict ordering constraint, enabling relational modeling while maintaining a well-defined directional flow of temporal information.

Graph Feature Representation Learning Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the aforementioned directed acyclic graph, where $\mathcal{V}$ refers the set of nodes which are connected by a set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Given a feature vector $z_i^t \in \mathbb{R}^D$ attached to the corresponding node, following their temporal ordering, we perform graph feature representation learning over the nodes using GraphSAGE [19], which follows the scheme of message passing. Specifically, nodes are updated by neighborhood feature aggregation, following the underlying graph structure [20].

2.2 Asymmetric Treatment Response Loss

Population-level Alignment We propose a population-level alignment objective that structures the embedding space according to the response status of a patient (Figure 1, bottom left). Let $z_i \in \mathbb{R}^D$ denote the embedding of a patient i, our goal is twofold: first, encourage similarity among responder patients (i.e. attract embeddings) and second, discourage similarity between responders and non-responders (i.e. repel embeddings), based on Cosine similarity, outlined as

$$\mathcal{L}_{\text{align}} = \underbrace{1 - \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \text{sim}(z_i, z_{\pi(i)})}_{\text{① attract}} + \underbrace{\frac{1}{|\mathcal{N}|} \sum_{(i,j) \in (\mathcal{N}, \mathcal{R})} \text{sim}(z_i, z_j)}_{\text{② repel}}. \tag{1}$$

① **attract** To promote a shared latent space among responder patients $\mathcal{R}$, we randomly permute the embeddings within the responder group and align each responder with another responder. Formally, let π be a random permutation over indices in $\mathcal{R}$, the responder attract term aims at increasing the pairwise similarity within the responder population, encouraging embeddings to occupy a compact region in latent space. Moreover, as the permutation π changes across iterations, the loss approximates a stochastic estimate of responder coherence rather than enforcing fixed pairwise matches within this subgroup.

② **repel** To promote discriminative separation between responders $\mathcal{R}$ and non-responders $\mathcal{N}$, we penalize similarity across these groups. Minimizing the repel term, pushes embeddings of the two subpopulations apart, thereby increasing inter-class separation. Furthermore, we intentionally omit the attract term for

non-responders, as their heterogeneous clinical outcomes such as stable disease, progressive disease, or partial response are associated with distinct underlying imaging patterns, making a shared latent representation ill-posed.

Patient-level Representation Decorrelation We propose a patient-level representation decorrelation objective, where the key idea is that embeddings from different timepoints of the same patient should not collapse to redundant representations, but rather encode complementary aspects of disease evolution. To this end, we penalize similarity between adjacent timepoint embeddings of responder patients $\mathcal{R}$ using Cosine similarity, formally defined as

$$\mathcal{L}_{\text{decorrelate}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \frac{1}{T-1} \sum_{t=0}^{T-2} \text{sim}\Big(z_i^{(t)}, z_i^{(t+1)}\Big) \,. \tag{2}$$

Minimizing the objective above encourages embeddings from different timepoints of the same patient to become decorrelated. As a consequence, each temporal segment $z_i^{(t)}$ of a patient trajectory is promoted to capture distinct, non-redundant factors of variation relevant for treatment response.

Patient-level Temporal Relationships We propose a patient-level temporal relationship loss, with the key idea of modeling how latent embeddings $\{z_i^t\}_{t=0}^{T-1}$ of patient i evolve over time. While the decorrelation term enforces complementary information across timepoints, we now constrain the latent trajectory to follow a consistent additive progression $\tilde{z}_i$, formally defined as

$$\mathcal{L}_{\text{temporal}} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \Big[1 - \text{sim}\Big(\tilde{z}_i, z_i^{(T)}\Big)\Big] \quad \text{where} \quad \tilde{z}_i = \sum_{t=0}^{T-2} \Big(z_i^{(t+1)} - z_i^{(t)}\Big) \,. \tag{3}$$

Minimizing the objective above encourages temporal consistency. The representation at the final timepoint T must be predictable from the sequence of intermediate latent transitions. Intuitively, the embeddings are encouraged to follow a coherent temporal trajectory rather than independent snapshots, thereby regularizing the model to encode meaningful disease progression dynamics.

Loss Function The overall *Asymmetric Treatment Response Loss* applies the introduced objectives asymmetrically to responders $\mathcal{R}$ and non-responders $\mathcal{N}$:

$$\mathcal{L}_{\text{total}}(i) = \mathbf{1}_{\{i \in \mathcal{R}\}} \big(\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{decorrelate}} + \mathcal{L}_{\text{temporal}}\big) + \mathbf{1}_{\{i \in \mathcal{N}\}} \mathcal{L}_{\text{repel}}$$

3 Experiments and Results

Data and Preprocessing In this work we use the publicly available ISPY2 dataset [12, 13] and select 585 patients (204 responders) with complete DCE-MRI acquisitions at four neoadjuvant chemotherapy (NACT) timepoints, comprising one pre-treatment scan and three consecutive examinations acquired during

therapy prior to surgery. Data preprocessing follows Jing et al. [8], resulting in a longitudinal sequence of four images per patient, each represented as a tensor of shape $(3, 64, 64, 64)$. The three channels encode early contrast enhancement (phase 1 minus phase 0), late contrast enhancement (phase 4 minus phase 0), and functional tumor volume (FTV), which captures metabolically active tumor regions defined via pharmacokinetic thresholds on DCE-MRI [21]. To ensure reproducibility and enable fair comparison, we release a user-friendly library at the Python Package Index (PyPI) implementing the full data pipeline.

Implementation Details All models are trained and evaluated using stratified 5-fold cross-validation (3/1/1 train/val/test split per fold). Optimization is performed with SGD (learning rate 10^{-2}, and batch size 16 with gradient accumulation over 4 batches) for 100 epochs using a cosine annealing scheduler. The model with the highest validation AUC is selected for testing. To ensure fair comparison, all methods share identical data splits, inputs, and the same ResNet18 feature encoder implementation [22], such that performance differences reflect only the respective spatio-temporal learning strategies. DINOv3 [23] is included as a strong training-free foundational model baseline, using the 2D axial volume slice containing the largest metabolically active tumor region (based on the FTV channel). All vision baselines employ the same three-layer MLP classification head, with the CNN+LSTM approach additionally incorporating an LSTM layer for sequential modeling [8]. Supervised representation learning methods are evaluated using a support vector classifier. Experiments are conducted on an NVIDIA RTX A6000 Ada GPU. Details can be found in our code repository.

3.1 Empirical Results and Experimental Analysis

Table 1 presents a comparison of our GNN-pCR framework against several supervised and supervised representation learning (SupRL) baselines for binary breast cancer pCR prediction. We evaluated [11] on the reference 2D data using 5-fold cross-validation and, in line with the authors [25], observed substantial performance variability, achieving an AUC of 0.5610. To ensure fair evauluation, we extended the method to 3D using our unified encoder, denoted as 3D-$\mathrm{L}_{\mathrm{ART}}$. Overall, our GNN-pCR consistently achieves the best performance across all evaluation metrics. In particular, compared to the strongest respective baseline model, our approach improves bACC from 0.6520 [11] to 0.6844, AUC from 0.6951 (best CNN baseline) to 0.7203, and Matthews correlation coefficient (MCC) from 0.2922 [11] to 0.3561. These results suggest that structured modeling of longitudinal relationships as proposed in our GNN-pCR approach, successfully translates into consistent improvements across all evaluation criteria and models. The performance improvement is even larger when compared to the SupRL approach of [24], the sequence modeling strategy based on CNN+LSTM [8] and the recent transformer vision foundation model DINOv3 [23].

Ablation Study We conduct an ablation study by systematically removing individual loss terms and the GNN module from the full model while keeping all

Table 1. Comparison of GNN-pCR with supervised and supervised representation learning (SupRL) baselines, with an ablation study of its key components. Performance is reported as the mean value over 5-folds, using balanced accuracy (bACC), Area Under the Receiver Operating Characteristic Curve (AUC) and Matthews correlation coefficient (MCC). **Best in bold**, second best underlined and ↑ means higher is better. Wilcoxon signed-rank test: $^{(*)}$refers to p-values ≤ 0.05, $^{(\dagger)}$refers to p-values ≤ 0.0625.

Method	Backbone	bACC ↑	AUC ↑	MCC ↑
Supervised baselines				
CNN + MLP	ResNet18	$0.6466^{(\dagger)}$ ± 0.03	$\underline{0.6951}^{(\dagger)}$ ± 0.03	$0.2864^{(\dagger)}$ ± 0.06
CNN + LSTM[8]	ResNet18	$0.6107^{(\dagger)}$ ± 0.05	0.6825 ± 0.06	$0.2278^{(\dagger)}$ ± 0.10
DINOv3[23] + MLP	ViT-L/16	$0.5628^{(*)}$ ± 0.04	$0.6020^{(*)}$ ± 0.03	$0.1316^{(*)}$ ± 0.09
SupRL baselines				
3D-L_{ART}[11]	ResNet18	$\underline{0.6520}^{(\dagger)}$ ± 0.03	0.6857 ± 0.02	$\underline{0.2922}^{(*)}$ ± 0.06
SSL-TOPC[24]	ResNet18	$0.6224^{(*)}$ ± 0.02	0.6775 ± 0.04	$0.2434^{(*)}$ ± 0.03
GNN-pCR (ours)	ResNet18	**0.6844** ± 0.01	**0.7203** ± 0.02	**0.3561** ± 0.02
Ablation Study				
w/o alignment loss	ResNet18	$0.6162^{(*)}$ ± 0.03	0.6936 ± 0.04	$0.2490^{(*)}$ ± 0.06
w/o temporal loss	ResNet18	$0.6589^{(*)}$ ± 0.03	0.7025 ± 0.06	$0.3047^{(*)}$ ± 0.05
w/o decorrelation loss	ResNet18	$0.6350^{(*)}$ ± 0.01	$0.6594^{(*)}$ ± 0.03	$0.2684^{(*)}$ ± 0.03
w/o GNN	ResNet18	0.6610 ± 0.02	0.6895 ± 0.04	$0.3077^{(\dagger)}$ ± 0.04

other components untouched. Results can be found in Table 1. Removing the alignment loss leads to a consistent performance drop across all metrics, with a substantial decline in e.g., bACC from 0.6844 to 0.6162. This finding is consistent with the objective of enhancing class separability between responders and non-responders, indicating that the alignment loss constitutes a key factor in shaping discriminative pCR representations. Similarly, removing the temporal loss ($0.6844 \rightarrow 0.6589$) or decorrelation loss ($0.6844 \rightarrow 0.6350$) leads to considerable performance declines across all metrics. These findings suggest that both encouraging complementary features (via decorrelation loss) and enforcing temporal consistency (via temporal loss) are essential to capture relevant longitudinal changes in responder patients. Finally, replacing the GNN with a linear prediction head also reduces performance ($0.6844 \rightarrow 0.6610$), underscoring that relational modeling of temporal interactions across timepoints provides benefits beyond those achievable with dense feature connections alone.

Early Response Prediction To assess the capability of the models to predict therapy response at early stages, we restrict the available longitudinal information to partial time series and evaluate performance for consecutive time intervals ($t_0 \rightarrow t_1$, $t_0 \rightarrow t_2$, and $t_0 \rightarrow t_3$). Results are summarized in Table 2. Across all temporal settings, our proposed GNN-pCR consistently achieves the strongest performance. Notably, the superiority of the full time-series model ($t_0 \rightarrow t_3$)

Table 2. Quantitative comparison of methods for (i) early response prediction and (ii) impact of inter-scan time differences. Performance is reported as the mean value over 5-folds, using balanced accuracy (bACC), for consecutive timepoints $t_0 \rightarrow t_1$, $t_0 \rightarrow t_2$ and $t_0 \rightarrow t_3$. **Best in bold**, second best underlined and ↑ means higher is better.

Method	Inter-Scan Time Diff.	bACC ↑		
		$t_0 \rightarrow t_1$	$t_0 \rightarrow t_2$	$t_0 \rightarrow t_3$
CNN + MLP	✗	0.5042 ± 0.03	0.6508 ± 0.03	0.6466 ± 0.03
	✓	0.5059 ± 0.02	0.6479 ± 0.03	0.6447 ± 0.02
CNN + LSTM[8, 16]	✗	0.5237 ± 0.05	0.5721 ± 0.03	0.6107 ± 0.05
	✓	0.5163 ± 0.04	0.5969 ± 0.03	0.6513 ± 0.02
3D-L_{ART}[11]	✗	0.5586 ± 0.05	0.6631 ± 0.03	0.6520 ± 0.03
	✓	0.5369 ± 0.06	0.6391 ± 0.05	0.6616 ± 0.02
GNN-pCR (ours)	✗	<u>0.5820</u> ± 0.04	**0.6823** ± 0.01	<u>0.6844</u> ± 0.01
	✓	**0.5937** ± 0.06	<u>0.6689</u> ± 0.04	**0.6853** ± 0.03

translates to earlier prediction scenarios. When only two timepoints are available ($t_0 \rightarrow t_1$), GNN-pCR attains the highest bACC (0.5937), clearly outperforming CNN-based and recurrent baselines. The margin over the second-best SupRL method (3D-L_{ART}) is particularly pronounced in this most challenging early setting, highlighting the benefit of our novel learning objectives combined with explicit modeling of relational structure under limited temporal data. We further analyze the impact of explicitly incorporating intra-patient time differences across scans as an additional feature. The effect is method-dependent. The CNN+LSTM [16] benefits most at the full time-series ($t_0 \rightarrow t_3$), improving from 0.6107 to 0.6513, although starting from a comparatively low score. In contrast, CNN, 3D-L_{ART} and GNN-pCR exhibit only moderate changes when adding time-gap information, which we attribute to the systematic design of the ISPY2 study, which shows near homogeneous time difference distribution across patients and their respective scans [12, 13].

4 Discussion and Conclusion

In this work, we have introduced a novel supervised framework for learning imaging-based patient treatment trajectories by combining a time-aware graph neural network with relational temporal modeling alongside three complementary trajectory-level objectives. To this end, the underlying graph structure explicitly encodes dependencies across timepoints, enabling structured representation learning aligned with clinically meaningful disease progression. Our experimental evaluations demonstrate consistent improvements over both supervised vision baselines and existing supervised representation learning approaches across multiple classification metrics. Ablation studies confirm the contribution

of each novel loss component and highlight the added value of relational modeling via the GNN. Notably, early response prediction results suggest that the learned representations capture clinically relevant longitudinal characteristics. Incorporating inter-scan time differences, although holding promise, yielded only moderate gains, likely due to the structured study design of ISPY2 and the resulting less pronounced variability in acquisition intervals across patients and scans. Even though our evaluations are based on breast cancer pCR prediction, our proposed framework is designed as a general approach for longitudinal SupRL. To this end, future work will evaluate the proposed framework across a broader range of longitudinal clinical prediction and monitoring tasks, potentially in combination with complementary clinical biomarkers.

Acknowledgments. Authors JK, SMF and JAS gratefully acknowledge financial funding received by Munich Center for Machine Learning. Authors JCP and JK gratefully acknowledge funding from the Wilhelm Sander Foundation in Cancer Research (2022.032.1). Author SMF has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 515279324/SPP 2177. JCP acknowledges funding from the Else Kröner-Fresenius Stiftung (2023-EKES.01). KL and RO acknowledge funding from EU Horizon Europe grant 101057699 (RadioVal). DML and JAS acknowledge funding from HELMHOLTZ IMAGING, a platform of the Helmholtz Information and Data Science Incubator. IJ acknowledges funding from Vienna Science and Technology Fund (WWTF, PREDICTOME [10.47379/LS20065]), 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. World Health Organization. Breast cancer: Global patterns of incidence, mortality, and survival. https://www.who.int/news-room/fact-sheets/detail/breast-cancer, 2026. Accessed: 2026-02-11.
2. Wen Li, David C Newitt, Jessica Gibbs, Lisa J Wilmes, Ella F Jones, Vignesh A Arasu, Fredrik Strand, Natsuko Onishi, Alex Anh-Tu Nguyen, John Kornak, Laura J Esserman, Don Berry, and Nola M Hylton. Predicting breast cancer response to neoadjuvant treatment using multi-feature MRI: results from the I-SPY 2 trail. *NPJ Breast Cancer*, 6(1):63, 2020.
3. Laura M Spring, Geoffrey Fell, Andrea Arfe, Chandni Sharma, Rachel Greenup, Kerry L Reynolds, Barbara L Smith, Brian Alexander, Beverly Moy, Steven J Isakoff, Giovanni Parmigiani, Lorenzo Trippa, and Aditya Bardia. Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: a comprehensive meta-analysis. *Clinical Cancer Research*, 26(12):2838–2848, 2020.
4. N.M. Hylton and et al. Locally advanced breast cancer: MR imaging for prediction of response to neoadjuvant chemotherapy-results from ACRIN 6657/ISPY trial. *Radiology*, 279(1):44–55, 2016.
5. Yunsong Peng et al. Pretreatment DCE-MRI-based deep learning outperforms radiomics analysis in predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer. *Frontiers in Oncology*, 12:846775, 2022.

6. Nabeeha Khan, Richard Adam, Pauline Huang, Takouhie Maldjian, and Tim Q Duong. Deep learning prediction of pathologic complete response in breast cancer using MRI and other clinical data: a systematic review. *Tomography*, 8(6):2784–2795, 2022.
7. Aaquib Syed, Richard Adam, Thomas Ren, Jinyu Lu, Takouhie Maldjian, and Tim Q Duong. Machine learning with textural analysis of longitudinal multiparametric MRI and molecular subtypes accurately predicts pathologic complete response in patients with invasive breast cancer. *PLOS One*, 18(1):e0280320, 2023.
8. Bowen Jing, Kai Wang, Erich Schmitz, Shanshan Tang, Yunxiang Li, You Zhang, and Jing Wang. Prediction of pathological complete response to chemotherapy for breast cancer using deep neural network with uncertainty quantification. *Medical Physics*, 51(12):9385–9393, 2024.
9. Wen Li, Savannah C Partridge, David C Newitt, Jon Steingrimsson, Helga S Marques, Patrick J Bolan, Michael Hirano, Benjamin Aaron Bearce, Jayashree Kalpathy-Cramer, Michael A Boss, et al. Breast multiparametric MRI for prediction of neoadjuvant chemotherapy response in breast cancer: the BMMR2 challenge. *Radiology: Imaging Cancer*, 6(1):e230033, 2024.
10. Eriseld Krasniqi, Lorena Filomeno, Teresa Arcuri, Gianluigi Ferretti, Simona Gasparro, Alberto Fulvi, Arianna Roselli, Loretta D'Onofrio, Laura Pizzuti, Maddalena Barba, et al. Multimodal deep learning for predicting neoadjuvant treatment outcomes in breast cancer: a systematic review. *Biology Direct*, 20(1):72, 2025.
11. Ivana Janíčková, Yen Y Tan, Thomas H Helbich, Konstantin Miloserdov, Zsuzsanna Bago-Horvath, Ulrike Heber, and Georg Langs. Temporal representation learning of phenotype trajectories for pCR prediction in breast cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 606–615. Springer, 2025.
12. Wen Li, DC Newitt, J Gibbs, LJ Wilmes, EF Jones, VA Arasu, F Strand, N Onishi, AAT Nguyen, J Kornak, et al. I-SPY 2 breast dynamic contrast enhanced MRI trial (ISPY2)(version 1). *The Cancer Imaging Archive (TCIA)*, 2022.
13. David C Newitt, Savannah C Partridge, Zheng Zhang, Jessica E Gibbs, Thomas L Chenevert, Mark A Rosen, Patrick J Bolan, Helga Marques, J Romanoff, L Cimino, and Nola M Hylton. ACRIN 6698/I-SPY2 breast DWI dataset (2021). *DOI: https://doi. org/10.7937/tcia. kk02-6d95*, 2025.
14. Maria Colomba Comes, Annarita Fanizzi, Samantha Bove, Vittorio Didonna, Sergio Diotaiuti, Daniele La Forgia, Agnese Latorre, Eugenio Martinelli, Arianna Mencattini, Annalisa Nardone, et al. Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs. *Scientific Reports*, 11(1):14123, 2021.
15. Song Zhang, Siyao Du, Caixia Sun, Bao Li, Lizhi Shao, Lina Zhang, Kun Wang, Zhenyu Liu, and Jie Tian. M2fusion: Multi-time multimodal fusion for prediction of pathological complete response in breast cancer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 458–468. Springer, 2024.
16. Ruggiero Santeramo, Samuel Withey, and Giovanni Montana. Longitudinal detection of radiological abnormalities with time-modulated LSTM. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 326–333. Springer, 2018.
17. Jessica Kächele, Dimitrios Bounias, Alexandra Ertl, and Klaus Maier-Hein. On tackling domain shift in breast MRI using only publicly-available data: Reproducible breast cancer segmentation and pCR prediction. In *Deep Breast Workshop*

on AI and Imaging for Diagnostic and Treatment Challenges in Breast Care, pages 310–319. Springer, 2025.
18. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
19. Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems*, 2017.
20. Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.
21. David C Newitt, Sheye O Aliu, Neil Witcomb, Gal Sela, John Kornak, Laura Esserman, and Nola M Hylton. Real-time measurement of functional tumor volume by MRI to assess treatment response in breast cancer neoadjuvant clinical trials: validation of the aegis SER software platform. *Translational Oncology*, 7(1):94–100, 2014.
22. M Jorge Cardoso, Wenqi Li, and Richard Nic Brown. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.
23. Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.
24. Emily Kaczmarek, Justin Szeto, Brennan Nichyporuk, and Tal Arbel. SSL-AD: Spatiotemporal self-supervised learning for generalizability and adaptability across alzheimer's prediction tasks and datasets. *arXiv preprint arXiv:2509.10453*, 2025.
25. Janíčková I. Temporal representation learning of phenotype trajectories for pCR prediction in breast cancer. https://github.com/cirmuw/temporal-representation-learning/blob/a1c721872bb6c2caa4dbe7c352cd878e283e3dc8/README.md, 2026. GitHub repository, commit a1c7218, accessed 25 Feb 2026.