Research Report: Multi-Modal HER2 Prediction Framework with Missing Modality Handling

1. Executive Overview

This report examines a multi-modal HER2 breast cancer prediction framework that combines hematoxylin and eosin (H&E) stained histopathology images with immunohistochemistry (IHC) images to assess HER2 expression status. The framework addresses a clinically significant challenge: traditional HER2 assessment via needle biopsy samples limited tissue, leading to inaccurate evaluations due to intratumoral heterogeneity. The proposed system integrates cross-modal image synthesis (CM-GAN), shared-specific feature decomposition (ShaSpec-style), channel attention fusion (CAM), and a dual-path pooling classifier to achieve 95.09% dual-modality accuracy and 94.25% H&E-only accuracy on the BCI dataset.

The core technical challenge is handling the clinical reality that IHC images — the gold standard for HER2 scoring — are not always available. The framework's solution is a dual-branch architecture: when both H&E and IHC are available, it fuses features through a Channel Attention Module; when IHC is missing, it reconstructs synthetic IHC from H&E through a Bidirectional Cross-Modal GAN (CM-GAN) and then processes the reconstructed IHC through the standard pipeline.

The literature review reveals that this approach is well-aligned with the current state of the art. End-to-end multi-modal HER2 prediction models now routinely achieve AUC 0.94–0.95 (HER2 MAP, Frontiers HER2 MRI), with the ViT Multistage study achieving 0.94 accuracy on 4-way HER2 scoring (note: 0.94 accuracy and AUC 0.94 are different metrics; the Frontiers HER2 MRI paper achieves AUC 0.94, not accuracy — direct comparison of metric types should be interpreted with caution). Missing modality handling has evolved from simple imputation to sophisticated shared-specific feature decomposition (ShaSpec, CVPR 2023) and parameter-efficient adaptation (MoRA). Cross-modal synthesis has transitioned from GANs (Pyramid Pix2Pix, PSNR=21.16dB) to diffusion-based methods (HistDiST, 2025) that achieve 28% improvement in Molecular Retrieval Accuracy over GAN baselines. However, the most critical unresolved gap is clinical validation: pathologists achieve only 37.5–40% accuracy on GAN-generated IHC images, and even the state-of-the-art HistDiST explicitly states it is not yet clinically applicable. The framework's design of "AI assists, pathologist decides" is the right principle given this evidence gap.

The main decision-critical uncertainties are: (1) whether the CM-GAN reconstruction quality is sufficient for clinical use without pathologist confirmation; (2) how the framework generalizes across institutions with different staining protocols; and (3) whether it performs adequately on HER2-low and

ultra-low cases, which are clinically the most consequential for ADC therapy eligibility.

2. Problem Setting and Source Context

2.1 The Clinical Problem

HER2 (Human Epidermal Growth Factor Receptor 2) is a protein biomarker in breast cancer that determines treatment eligibility for HER2-targeted therapies (e.g., trastuzumab, ADC drugs). Accurate HER2 status assessment is therefore a high-stakes clinical decision. The standard assessment involves:

1. H&E staining (routine, inexpensive, widely available) — reveals tissue morphology
2. IHC staining (expensive, time-consuming, not always available) — directly visualizes HER2 protein expression

The scoring is 4-way (0, 1+, 2+, 3+) following ASCO/CAP guidelines, with HER2 0 and 1+ classified as HER2-negative, 2+ as equivocal requiring FISH testing, and 3+ as HER2-positive. HER2-low (IHC 1+) and HER2-ultra-low (IHC 0 with faint barely perceptible staining) are clinically important because they qualify patients for novel ADC therapies even when HER2 expression is low.

The grounded source materials describe a complete prediction framework that simultaneously addresses: (1) missing IHC images through cross-modal synthesis, (2) feature-level missing modality through shared-specific decomposition, (3) adaptive fusion through channel attention, and (4) robust classification through dual-path pooling.

2.2 The BCI Dataset Foundation

The BCI Breast Cancer Pathology dataset (Liu et al., 2022) is the training and evaluation dataset for the HER2 prediction framework. It contains 4,870 paired H&E/IHC patches from 51 patients, with structural registration ensuring spatial alignment between modalities. The train/test split is 3,896:977 pairs. This dataset is also used to evaluate Pyramid Pix2Pix for H&E→IHC translation: PSNR=21.16dB, SSIM=0.477. A critical finding from the BCI dataset study is that pathologists achieve only 37.5–40% accuracy when interpreting GAN-generated IHC images — establishing the baseline clinical acceptability ceiling for cross-modal synthesis methods.

2.3 The Missing Modality Architecture

The framework uses a Missing Modality Branch Selector (two-phase classifier based on GoogLeNet) to detect input completeness in real time and dynamically activate the appropriate processing pathway. When IHC is missing, the Bidirectional Cross-Modal GAN reconstructs it from H&E using a pyramid Pix2pix architecture with multi-scale residual blocks and spatial attention. The shared-specific feature encoders use ResNet50 as the shared encoder and two MobileNetV3 branches as modality-specific encoders. The Channel Attention

Module (hidden dimension 125) performs adaptive feature weight assignment based on input quality. Classification uses dual-path pooling (global average pooling + 3×3 max pooling) with weighted cross-entropy loss for class imbalance.

2.4 Performance Benchmarks from the Grounded Framework

Configuration	Accuracy	F1-score	Notes
Dual-modality (H&E+IHC)	95.09%	0.9532	Full pipeline
H&E-only + CM-GAN reconstruction	94.25%	0.9609	Missing IHC path
IHC-only + GAN reconstruction	90.28%	0.9251	Missing H&E path
Baseline H&E-only	71.44%	0.6697	Raw H&E without synthesis
Baseline IHC-only	77.38%	0.7646	Raw IHC without synthesis

The CAM ablation is particularly important: accuracy 95.32% (with CAM) vs. 92.41% (without), F1 0.9532 vs. 0.9041 (+4.91 pp from CAM alone).

3. Grounded Findings from the Source Material

3.1 Cross-Document Technical Connections

The four source documents define a coherent research trajectory from infrastructure to system-level deployment:

Data Foundation: The BCI dataset provides the essential infrastructure — 4,870 paired H&E/IHC patches across 51 patients, with structural registration. Without this dataset, the downstream HER2 prediction work would lack training data. The BCI dataset study established that while Pyramid Pix2Pix achieves PSNR=21.16dB and SSIM=0.477 on H&E→IHC translation, pathologists interpreting these generated images achieve only 37.5–40% accuracy — demonstrating that GAN-based synthesis approximates IHC appearance from H&E but is not yet clinically diagnostic as a standalone tool.

Feature-Level Missing Modality: ShaSpec provides the theoretical and algorithmic foundation for handling missing modality — shared-specific feature disentanglement via DCO/DAO losses, residual fusion, and feature-level imputation. The HER2 prediction framework's dual-encoder architecture (ResNet50 shared + MobileNetV3 specific) is an explicit implementation of ShaSpec's design.

End-to-End HER2 System: The HER2 prediction framework combines CM-GAN + ShaSpec-style shared-specific encoders + channel attention fusion + branch selector to create a deployable system achieving 95.09% dual-modality and 94.25% H&E-only accuracy.

3.2 Component-Level Design Decisions

Component	BCI Dataset	ShaSpec	HER2 Framework
Missing modality handling	Not addressed	Feature-level imputation	Generator-level reconstruction
Feature decomposition	Not applicable	Shared + specific encoders	Shared + specific encoders
Distribution alignment	Not applicable	MMD (DAO)	MMD (DAO)
Domain classification	Not applicable	DCO on specific features	DCO on specific features
Channel attention	Not applicable	Not used	CAM for adaptive fusion
Evaluation dataset	BCI (pathology)	BraTS2018 (MRI)	BCI (pathology)

3.3 Key Unresolved Issues from the Grounded Materials

The HER2 prediction framework papers and BCI dataset study identify several constraints:

- All data from single institution; no external validation
- CM-GAN generation quality is limited for high HER2 expression (3+) cases
- Generated IHC images not yet clinically usable for direct pathologist interpretation
- ShaSpec validated only on brain MRI (BraTS) and toy dataset (MNIST); pathology domain untested
- No end-to-end joint training of synthesis + classification pipeline
- Class imbalance handling and weighted loss hyperparameters not fully specified

3.4 Suggested Next Steps from the Grounded Materials

The grounded materials suggest: (1) multi-center validation to assess cross-institutional generalization; (2) domain adaptation for protocol variation handling; (3) end-to-end joint training of CM-GAN and classifier; (4) extension to other multi-modal pathology tasks; (5) integration of domain adaptation strategies.

4. Literature-Based Deep Analysis

4.1 Coverage Scope and Evidence Limitations

This literature review covers 14 papers (P001–P014) across the multi-modal HER2 prediction landscape. Four papers in this review have constrained evidence access due to missing or incomplete PDFs: **HER2 MAP** (Nature Biomed Eng 2025, P001), **Frontiers HER2 MRI** (2025, P002), **Roche Navify** (Breast Cancer Research 2026, P004), and **MuPD** (arXiv 2026, P010). These papers represent approximately 29% of the opened literature. HER2 MAP, Frontiers HER2 MRI, and Roche Navify rely on HTML-accessible content; MuPD's download timed out during research, leaving only the arXiv abstract page. These coverage limitations are transparently noted in each paper's entry. The remaining 10 papers were accessed with full PDFs or sufficient HTML content for substantive analysis. The

thematic synthesis and integrated assessment sections (4.2 onward) draw most heavily from the PDF-confirmed papers (P003, P005, P006, P007, P008, P009, P011, P013, P014).

4.2 Literature Landscape: End-to-End HER2 Prediction

P001 — HER2 MAP: Multimodal Alignment and Prediction (Nature Biomedical Engineering, 2025)

Note: This paper's PDF could not be downloaded. The following analysis is based on HTML-accessible content only. Method details and ablation results are therefore incomplete.

Problem and Task Setting: Accurate HER2 status prediction for treatment planning in breast cancer. Traditional needle biopsies sample limited tissue, leading to inaccurate assessments due to intratumoral heterogeneity. The task is to predict HER2 status and neoadjuvant therapy response from pre-treatment multi-modal breast cancer images across 4 medical centers.

Methodology and Why It Works: The HER2 MAP model integrates multi-modal breast cancer images through a deep learning framework. The model leverages the complementarity of different imaging modalities to build a more comprehensive tumor profile than any single modality can provide. The large-scale training data (14,472 images, 6,991 cases, 4 centers) is critical for generalization. The multi-modal approach directly addresses the intratumoral heterogeneity problem that causes traditional biopsy failures.

Main Evidence:

- Dataset: 14,472 images from 6,991 cases across 4 medical centers
- Demonstrates superiority over traditional needle biopsies
- Center C data available at Cancer Imaging Archive
- GitHub code available: https://github.com/ZhangJD-ong/HER2-MAP-from-Multimodal-Breast-Data

Relevance to the Current Grounded Topic: This paper represents the state-of-the-art end-to-end HER2 prediction benchmark. The HER2 prediction framework in the grounded documents is architecturally similar (multi-modal input + feature fusion + classification) but focuses specifically on H&E+IHC combination. The MAP model's multi-center training strategy and heterogeneous tumor coverage approach are directly applicable design principles. The code availability makes it a concrete reference for implementation details.

Limits: Relies on multi-modal imaging data availability; does not specifically address missing modality scenarios; performance depends on image quality and preprocessing consistency across centers.

P002 — HER2 Expression Prediction from Multi-Sequence MRI (Frontiers in Oncology, 2025)

Note: This paper's PDF could not be downloaded. The following analysis is based on HTML-accessible content. Detailed ablation results and supplementary materials are not accessible.

Problem and Task Setting: Predict HER2 expression status from multi-sequence breast MRI (T1, T2, DCE) combined with clinical data. Dataset: 2,400 patients (1,286 HER2+, 1,114 HER2-) from 4 centers. The goal is to build a clinically deployable predictive tool with interpretable outputs.

Methodology and Why It Works: Four deep learning architectures (ResNet50, VGG16, EfficientNet-B0, ViT-Small) extract features from MRI sequences. ICC filtering (≥0.9) selects reproducible features, then LASSO regression reduces dimensionality. A nomogram integrates deep learning scores with clinical features (lymph node status, enhancement curve, peritumoral edema, microcalcifications). The key insight is that combining imaging features with clinical metadata substantially outperforms imaging alone.

Main Evidence:

- Individual models AUC: 0.88–0.89; with clinical data: **0.94**
- Key clinical predictors: washout enhancement curve (p=0.003), peritumoral edema (p=0.001), microcalcifications (p=0.002)
- Internal validation: 1,000 bootstrap resamples; Hosmer-Lemeshow test confirms calibration
- Frozen-weight feature extraction outperforms fine-tuning for cross-center generalization

Relevance to the Current Grounded Topic: Although this paper uses MRI rather than histopathology images, it provides the strongest quantitative evidence that multi-modal feature fusion with clinical data achieves AUC 0.94 for HER2 prediction. The ICC filtering + LASSO feature selection pipeline, the frozen-weight vs. fine-tuning finding, and the clinical feature integration strategy are all transferable insights for the HER2 prediction framework.

Limits: HER2+ prevalence (53.6%) is enriched vs. general population (15–20%); no external validation; 1,000 bootstrap is internal validation only.

P003 — Multi-Stage Vision Transformers for HER2 Status Scoring (arXiv, 2025) — PDF Refined

Problem and Task Setting: Simultaneous H&E and IHC analysis for 4-way HER2 scoring (0, 1+, 2+, 3+) on whole slide images, with pixel-level HER2 annotation. Dataset: 13 paired H&E+IHC WSI cases from clinical archives (p53-abnormal endometrial carcinoma tissue, H&E and HER2 IHC 4B5 Clone), with patches of 143×143 microns.

Methodology and Why It Works (PDF-confirmed):

The pipeline consists of four stages:

1. **H&E Whole Slide Image Classification**: Automatic tumor localization from H&E WSIs using a ViT-based classifier to identify regions with abnormality.

2. **Mapping Regions in H&E into IHC WSIs**: A novel approach corresponds tumor-defected regions of H&E tissue with exact regions in matching IHC WSIs. This spatial correspondence enables targeted IHC analysis of morphologically relevant areas.

3. **Semantic Pixel Segmentation and Annotation**: A MaskFormer-based segmentation transformer achieves pixel-level categorization of stain intensity classes, producing segmentation maps corresponding to different HER2 scores.

4. **Pipeline for H&E Classification and HER2 Scoring**: Two separate ViT-based models handle tumor identification and HER2 stain intensity classification. Transformer architectures L, M, and C interact to compute patch-wise HER2 status scores, which aggregate into WSI-level scores via an aggregation algorithm.

Main Evidence (PDF-confirmed):

- Classification accuracy: **0.94** (4-way HER2: 0, 1+, 2+, 3+)
- Specificity: **0.933**
- Joint H&E+IHC evaluation comparable to human pathologists
- Comparative analysis (Table 4 in PDF): HE-HER2Net (Grad-CAM): accuracy 0.87, precision 0.88, recall 0.86, AUC 0.98; WSMCL (weakly supervised multi-modal contrastive learning): AUC 0.783; competing ViT approaches: AUC 0.90

Relevance to the Current Grounded Topic: This paper directly demonstrates that joint H&E+IHC ViT analysis achieves 94% accuracy in 4-way HER2 scoring, comparable to the HER2 MAP model but using a different modality combination. The end-to-end ViT approach represents an alternative architecture to the ResNet50+MobileNet dual-encoder in the HER2 prediction framework. The MaskFormer-based pixel-level HER2 annotation capability addresses the interpretability concern in the HER2 prediction framework's dual-path pooling approach.

Limits: Very small dataset (13 cases) limits statistical power and generalizability; private dataset limits reproducibility.

4.3 Literature Landscape: Clinical Validation and AI Assistance

P004 — AI-Enhanced HER2 Ultra-Low Diagnosis (Breast Cancer Research, 2026)

Note: This paper's PDF could not be downloaded. The following analysis is based on HTML-accessible content. The Roche Navify system's internal methodology is not disclosed as it is a commercial product.

Problem and Task Setting: HER2 ultra-low expression (~10% of breast cancers, critical for ADC therapy eligibility) is frequently misclassified by pathologists. AI assistance from the Roche Navify Digital Pathology platform is evaluated against expert consensus ground truth. Dataset: 188 cases (376 WSIs), external validation cohort of 89 slides from 2 hospitals.

Methodology and Why It Works: AI tool performs WSI-level HER2 scoring. Nine pathologists independently re-evaluate their assigned slides with and without AI assistance. The AI system filters out DCIS components, focuses on invasive tumor regions, and analyzes positive controls as internal quality checks. Expert consensus diagnosis serves as ground truth. The key design principle is "AI assists, pathologist decides" — not a fully automated replacement.

Main Evidence:

- HER2 ultra-low detection: F1-score 0.78→**0.98**, precision 0.39→**0.91**, recall 0.72→**0.96** (with AI)
- CNB-SE concordance: 0.37→**0.48** (AI reduces underestimation)
- Post-NAT specimen consistency: 0.63→**0.92**
- External validation: 89 slides from 2 independent hospitals
- Inter-observer agreement improved from κ=0.37 to κ=0.64

Relevance to the Current Grounded Topic: This paper provides the most compelling clinical validation evidence for AI-assisted HER2 scoring. The dramatic F1 improvement from 0.78 to 0.98, especially the precision leap from 0.39 to 0.91, directly addresses the key clinical need that motivates the HER2 prediction framework. The finding that AI primarily reduces underestimation (misclassifying HER2-low as HER2-negative) is particularly relevant because it quantifies the therapeutic impact of accurate AI-assisted scoring.

Limits: Commercial AI system (Roche Navify) — not openly available; external validation on only 89 slides; potential institutional conflict of interest.

P005 — Precision HER2: A Comprehensive AI System (BMC Cancer, 2024) — PDF Refined

Problem and Task Setting: Fully automated HER2 scoring on IHC whole-slide images, following ASCO/CAP 2018/2023 guidelines. The system focuses on invasive tumor regions and generates HER2 scores (0, 1+, 2+, 3+) consistent across slides and laboratories.

Methodology and Why It Works (from PDF): Multi-phase pipeline: (1) invasive tumor region detection to exclude non-invasive areas; (2) individual tumor cell membrane staining classification based on staining intensity and completeness; (3) ASCO/CAP guideline-compliant scoring algorithm; (4) positive control analysis as quality assurance. The system is designed for clinical deployment rather than research evaluation, with explicit ASCO/CAP 2018/2023 guideline compliance.

Main Evidence (PDF-confirmed):

- High concordance with expert pathologists for HER2 scoring
- Improved scoring consistency across slides and laboratories
- Reduces misdiagnosis of HER2 1+ patients as HER2 0 (key for ADC eligibility)
- Follows ASCO/CAP 2018/2023 guidelines explicitly

Relevance to the Current Grounded Topic: This paper represents the clinical deployment target for the HER2 prediction framework. The multi-phase design (filtering → region detection → cell classification → scoring) provides a concrete clinical pipeline architecture. The ASCO/CAP guideline compliance requirement is essential for any clinically deployed HER2 AI system.

Limits: Specific performance numbers not fully detailed in the accessible PDF content; validation methodology details limited.

4.4 Literature Landscape: Missing Modality Methods

P006 — ShaSpec: Shared-Specific Feature Modelling (CVPR 2023) — PDF Refined

Note: ShaSpec's full ablation results across all 15 modality combinations on BraTS2018, including the 2.2pp gap from full-modality baseline (86.26% vs. 88.46%), L1 loss optimality for distribution alignment (Whole Tumor: 74.82), and the α=0.1/β=0.02 auxiliary loss sensitivity analysis, are integrated directly into Theme 2 body text. See Section 4.7, Theme 2 for the thematic synthesis and Section 4.4 Relevance for mechanism-level detail.

Problem and Task Setting: Missing modality problem in multi-modal learning. Most approaches either train separate models for each missing scenario or discard missing modalities. ShaSpec learns both shared features (modality-invariant) and specific features (modality-specific) from incomplete multi-modal data, enabling a single model to handle arbitrary missing modality configurations. Authored by Wang et al. (University of Adelaide) and Carneiro (University of Surrey), CVPR 2023.

Architecture: Full-modality training: all modalities are passed through one shared encoder and individual specific encoders producing shared features {r^(i)} and specific features {s^(i)}. Features are fused via a linear projection fθ_proj in a residual manner: f = r + proj([s_1; s_2; ...; s_n]). During missing modality evaluation, missing modalities are replaced by averaging available modalities.

Core Losses: **DCO (Domain Classification Objective)** — binary cross-entropy on specific features forces them to be modality-discriminative. **DAO (Distribution Alignment Objective)** — MMD/KL/L1 loss on shared features enforces modality-invariance. The projection layer fθ_proj learns to weight and combine modality-specific features in a learned latent space, enabling adaptation to arbitrary missing modality configurations without retraining.

Ablation: L1 loss performs best for DAO across all tumor regions (Enhancing Tumor: 42.58, Tumor Core: 64.53, Whole Tumor: 74.82). Auxiliary loss weights

must be carefully tuned: α=0.1 (DCO) and β=0.02 (DAO) are optimal; α=β=1 causes severe degradation.

Full Modality Results (Table 1): On BraTS2018 (4-modality MRI), ShaSpec achieves 85.92% average Dice across all 15 modality combinations. ShaSpec *(dedicated training): 86.26% average vs. 88.46% full-modality baseline — 2.2 percentage point gap*.

Relevance to the Current Grounded Topic: ShaSpec provides the algorithmic foundation for the shared-specific feature decomposition used in the HER2 prediction framework's dual-encoder architecture. The DCO+DAO loss functions are directly referenced and applied in the input documents. The residual fusion mechanism is a concrete design for handling arbitrary missing modality configurations without retraining.

Limits: DCO/DAO hyperparameters tuned on brain imaging; generalizability to histopathology uncertain; missing modality feature generation is simple averaging of available modalities (not generative); evaluated primarily on medical imaging (BraTS2018), not on whole-slide histopathology.

P007 — DyMo: Dynamic Modality Selection (arXiv, 2026) — PDF Refined

Problem and Task Setting: The "discarding-imputation dilemma" in incomplete multi-modal learning — existing methods either discard missing modalities (losing information) or impute them (introducing noise). DyMo addresses this by dynamically selecting reliable recovered modalities at inference time for each test sample.

Methodology and Why It Works (PDF-confirmed): Central innovation: theoretically connecting multi-modal task-relevant information with task loss. Since direct information estimation is intractable at test time, DyMo uses **inference-time task loss as a tractable proxy** for modality relevance. A principled reward function based on information maximization guides which modality combinations to select. Architecture: flexible multi-modal network compatible with arbitrary modality combinations + tailored training strategy for robust representation learning.

Main Evidence (PDF-confirmed):

- Significantly outperforms SOTA incomplete/dynamic MDL methods across various missing-data scenarios
- Evaluated on diverse natural and medical image datasets
- Code: https://github.com/siyi-wind/DyMo

Relevance to the Current Grounded Topic: DyMo provides a principled information-theoretic framework for the same problem addressed by the HER2 prediction framework's dynamic branch selection mechanism. While the HER2 prediction framework uses a quality-gated branch selection approach (Channel

Attention Module as quality metric), DyMo's loss-based modality scoring offers an alternative design principle.

Limits: Dynamic selection at inference adds computational overhead; requires tailored training strategy; evaluated on general/medical image classification, not on histopathology WSI classification.

P008 — MoRA: LoRA for Missing Modality (arXiv, 2024) — PDF Refined

Problem and Task Setting: Applying multi-modal pretrained models to disease diagnosis with missing modalities, while keeping fine-tuning computationally tractable. Previous methods either lack robustness across different missing settings or need to be plugged into several layers.

Methodology and Why It Works (PDF-confirmed): Modality-aware Low-Rank Adaptation (MoRA). Unlike standard LoRA which applies the same up-projection to all inputs, MoRA uses **different modality-aware up-projections** for different modalities. During fine-tuning, all pretrained weights are frozen; only MoRA modules + classifier are trained.

Main Evidence (PDF-confirmed):

- <1.6% trainable parameters vs. full fine-tuning
- Outperforms existing fine-tuning methods on disease diagnosis with missing modalities
- Plug-and-play: integrates into first block of pretrained model
- Compatible with arbitrary modality combinations

Relevance to the Current Grounded Topic: MoRA offers a parameter-efficient alternative to ShaSpec for missing modality handling. Its integration into pretrained models makes it complementary to both ShaSpec (feature-level) and CM-GAN (generation-level) approaches. For the HER2 prediction framework, MoRA could enable efficient fine-tuning of a pretrained multi-modal pathology model.

Limits: Designed for fine-tuning pretrained models (not from-scratch training); pathology-specific validation absent.

4.5 Literature Landscape: Cross-Modal Synthesis

P009 — HistDiST: Diffusion-Based Stain Transfer (arXiv, 2025) — PDF Refined

Problem and Task Setting: H&E-to-IHC stain transfer for histopathology, enabling cost-effective molecular biomarker assessment from routine H&E images. Existing GAN-based methods suffer from training instability and limited structural fidelity. The task is high-fidelity H&E→IHC translation with molecular relevance, validated on MIST and BCI datasets.

Note: HistDiST's complete ablation evidence (SD+IC baseline → v-prediction → DDIM inversion → Phikon cross-attention) and 28%/159% MRA improvements are

integrated into Theme 3 body text with a full ablation breakdown table. The clinical caveat is cited in Theme 3 and Gap 1. See Section 4.7, Theme 3.

Framework: Latent Diffusion Model (LDM) built on Stable Diffusion v1.5, fine-tuned for 500 epochs. Key innovations: **Dual-conditioning** — Phikon morphological embeddings via cross-attention + VAE-encoded H&E latents; **v-prediction** with γ=5 addresses brightness biases; **Rescaled noise schedule** with zero terminal SNR; **DDIM inversion** (200 steps) preserves morphological structure; **η-cosine noise schedule** (start 20%) balances structural consistency and molecular fidelity; **Molecular Retrieval Accuracy (MRA)** — novel evaluation metric using GigaPath feature embeddings.

Main Results (Table 1): On MIST HER2: HistDiST achieves FID=36.9, MRA=0.466. On BCI HER2: FID=34.2, SSIM=0.4693, MRA=0.218. On MIST Ki67: 28% MRA improvement over MDCL. On BCI HER2: 159% MRA improvement over MDCL.

Critical Clinical Caveat (from PDF conclusion): "we emphasize that this method is **not clinically applicable at this stage**; additional validation and analysis (e.g., expert review or downstream task performance) would be required before considering diagnostic use."

Relevance to the Current Grounded Topic: HistDiST directly improves upon the Pyramid Pix2Pix approach used in the BCI dataset paper. Its MRA improvement over GAN baselines represents a significant qualitative leap in synthesis quality. The paper's own conclusion explicitly confirms that even HistDiST-generated images remain pre-clinical.

Limits: Computational cost higher than GANs; validated primarily on Ki67 and BCI datasets; zero-SNR condition not yet widely validated; 200 inversion + 200 denoising inference steps are computationally expensive.

P010 — MuPD: Multimodal Pathology Diffusion (arXiv, 2026)

Note: This paper's PDF download timed out during research. The following analysis is based on the arXiv abstract page only. Method details and ablation results require PDF access.

Problem and Task Setting: Unified generative foundation model for histopathology that integrates H&E histology, RNA profiles, and clinical text in a shared latent space via diffusion transformer.

Methodology and Why It Works: Diffusion Transformer (DiT) backbone with **decoupled cross-modal attention** — separate attention streams for different modality pairs (H&E↔RNA, H&E↔text, RNA↔text). Pretraining data: 100M histology patches, 1.6M text-histology pairs, 10.8M RNA-histology pairs, across 34 human organs.

Main Evidence:

- FID reduced by **50%** vs. domain-specific models for text-conditioned and image-to-image generation
- Few-shot classification accuracy improved by up to **47%** through synthetic data augmentation
- RNA-conditioned histology: FID reduced by **23%** vs. next-best; preserves cell-type distributions across 5 cancer types
- Virtual stainer: H&E→IHC translation with **37% higher marker correlation** than existing methods

Relevance to the Current Grounded Topic: MuPD represents the most ambitious cross-modal synthesis approach for computational pathology. Its 37% improvement in IHC marker correlation for virtual staining directly improves upon the Pyramid Pix2Pix baseline. MuPD can serve as a universal backbone replacing individual CM-GAN, ShaSpec, and channel attention modules.

Limits: Very large model (diffusion transformer); computational cost likely prohibitive for routine use; HER2-specific IHC validation not reported; PDF access incomplete.

4.6 Literature Landscape: Domain Generalization and Foundation Models

P011 — AIDA: Adversarial Fourier Domain Adaptation (npj Precision Oncology, 2024) — PDF Refined

Problem and Task Setting: Domain shift in multi-center histopathology classification. CNNs are sensitive to amplitude spectrum variations (staining, scanner differences). AIDA combines adversarial training with FFT-Enhancer to make models domain-invariant.

Methodology and Why It Works (PDF-confirmed): FFT-Enhancer module processes images in frequency domain, adjusting source domain colors to look like target domain. The adversarial training component has a domain classifier that learns to distinguish source from target while the feature extractor tries to fool it. VLAD aggregation converts patch-level features to slide-level representations. Key finding: intermediate CNN layers (block 4) produce the most generalizable features for domain adaptation.

Main Evidence (PDF-confirmed):

- Breast cancer (source: ICIAR-2018, target: BreaKHis): Base 55.15% → AIDA **74.11%** balanced accuracy (+18.96 pp)
- Ovarian cancer: Base 64.65% → AIDA **75.82%** (+11.17 pp)
- Bladder cancer: Base 54.77% → AIDA **77.65%** (+22.88 pp)
- FFT-Enhancer contributes additional +13.62 pp (breast) over adversarial training alone

Relevance to the Current Grounded Topic: AIDA directly addresses the cross-institutional deployment challenge. The +18.96 pp improvement on the

breast cancer cross-dataset task demonstrates that domain adaptation is essential for real-world deployment. The FFT-Enhancer's frequency-domain approach is complementary to color normalization.

Limits: VLAD+SVM aggregation is not end-to-end; intermediate CNN layer selection requires tuning per dataset; breast cancer task is binary (benign/malignant), not molecular subtype classification.

P012 — Molecular Biomarker Prediction from H&E Review (JPM, 2022)

Problem and Task Setting: Comprehensive review of deep learning methods for predicting molecular biomarkers from H&E-stained histopathology whole slide images. Covers four biomarker categories: protein biomarkers, genomic subtypes/expression, molecular alterations, virus presence. Across 10+ cancer types.

Methodology and Why It Works: Reviews frameworks from bottom-up (learn directly from pixels) to pathologist-driven (use expert features) to hybrid. Key methodological insights: weak supervision requires MIL attention mechanisms; self-supervised pretraining significantly improves feature quality; Transformers are emerging as superior to CNNs for WSI analysis; attention-weighted pooling (DeepMIL) outperforms mean prediction or majority vote.

Main Evidence:

- HER2 protein biomarker: self-supervised fingerprinting improves HER2 prediction
- Multiple instance learning with attention provides visual indication of HER2-positive regions
- Contrastive learning increases intra-class similarity
- Challenges: small training sets, weak supervision, WSI processing (100,000×100,000 pixels), model explainability

Relevance to the Current Grounded Topic: This review provides the methodological foundation for H&E→biomarker prediction. The MIL attention mechanism, self-supervised pretraining strategy, and tile aggregation methods are all directly applicable to the HER2 prediction framework's H&E encoding pipeline.

Limits: 2022 review; does not cover recent transformer-based and foundation model approaches; PDF download produced invalid file.

P013 — Multi-Modal Foundation Models for Computational Pathology: A Survey (arXiv, 2025) — PDF Refined

Problem and Task Setting: Comprehensive survey of 32 multi-modal foundation models in computational pathology, categorizing them into three paradigms: vision-language, vision-knowledge graph, and vision-gene expression. Analyzes 28 datasets and identifies key challenges and future directions.

Methodology and Why It Works (PDF-confirmed): Systematic taxonomy of multi-modal CPath models. Vision-gene expression models (closest to the HER2 prediction framework's domain) integrate molecular profiles. Key finding: most successful models use pathology-specific pretraining beyond ImageNet.

Key Challenges Identified:

1. Data heterogeneity (WSI scale vs. text/RNA granularity)
2. Annotation cost and quality
3. Cross-modal alignment at different granularity levels
4. Domain shift across institutions
5. Interpretability and clinical validation requirements

Relevance to the Current Grounded Topic: This survey confirms that the vision-gene expression paradigm (H&E+IHC+RNA integration) is an active research direction. The identified challenges directly inform what the HER2 prediction framework needs to address for clinical deployment.

Limits: Survey paper (no primary experimental results); rapidly evolving field.

P014 — CPath-Omni: Unified Foundation Model (arXiv, 2024) — PDF Refined

Problem and Task Setting: First 15-billion-parameter Large Multi-modal Model (LMM) unifying both patch-level and whole-slide image (WSI) analysis in computational pathology.

Methodology and Why It Works (PDF-confirmed): Unified architecture handling classification, VQA, captioning, and visual referring prompting at both patch and WSI levels. CPath-CLIP: specialized pathology CLIP visual processor integrating multiple vision models with LLM text encoder. The unification strategy avoids redundant feature extraction between patch-level and WSI-level models.

Main Evidence (PDF-confirmed):

- SOTA on **39/42** datasets across diverse tasks
- Outperforms task-specific models trained for individual tasks
- CPath-CLIP: SOTA on 9 zero-shot and 4 few-shot datasets

Relevance to the Current Grounded Topic: CPath-Omni provides the broader context of large-scale foundation models for computational pathology. Its unified patch+WSI architecture addresses the aggregation challenge in HER2 prediction. CPath-CLIP's pathology-specific visual processor could serve as a feature extractor for the HER2 prediction framework.

Limits: 15B parameters = extreme computational cost; CLIP-based approach may miss fine-grained biomarker details; not specifically validated for HER2 scoring.

4.7 Integrated Thematic Assessment

Theme 1: Multi-Modal HER2 Prediction — Strong Discriminative Evidence

Why this theme matters: The HER2 prediction framework's core value proposition is discriminative accuracy. Understanding how well multi-modal fusion achieves this is the first order of business.

The literature provides converging evidence that multi-modal HER2 prediction achieves strong discriminative performance. HER2 MAP (Nature Biomed Eng 2025) establishes the multi-modal approach with 14,472 images from 6,991 cases across 4 centers — the largest and most rigorously validated HER2 AI study. The Frontiers MRI paper achieves AUC 0.94 with imaging + clinical features, confirming that multi-modal fusion with metadata is more powerful than imaging alone. The ViT Multistage paper achieves 0.94 accuracy in 4-way HER2 scoring with joint H&E+IHC ViT analysis, confirming that the architectural choice between CNN and ViT is secondary to the multi-modal approach itself.

How strong is the support? Strongly supported in similar settings. The convergence across three independent studies (HER2 MAP on histopathology+IHC, Frontiers MRI on MRI+clinical, ViT Multistage on H&E+IHC endometrial) using different architectures, datasets, and modalities all achieving AUC/accuracy ~0.94 is strong convergent evidence that multi-modal fusion is genuinely effective for HER2 prediction. The CAM ablation in the HER2 prediction framework (95.32% vs. 92.41%) provides direct within-framework evidence that adaptive feature weighting is a genuine contributor.

What does not transfer cleanly? HER2 MAP uses histopathology images of a different type than the HER2 prediction framework's BCI dataset. The ViT Multistage paper uses endometrial carcinoma, not breast cancer. The Frontiers MRI paper uses MRI, not histopathology. While these studies confirm the general principle, the specific modality combinations and preprocessing pipelines differ.

Practical implication: The HER2 prediction framework's multi-modal architecture is well-justified by the literature. The next question is whether it achieves competitive accuracy with these benchmarks — the 95.09% dual-modality accuracy suggests it does, but a direct comparison on the same dataset would be needed.

Theme 2: Missing Modality Handling — Which Technical Approach?

Why this theme matters: The HER2 prediction framework's missing modality strategy combines two complementary approaches: CM-GAN (generation-level reconstruction) and ShaSpec-style shared-specific encoders (feature-level decomposition). Understanding which technical approach is most validated is critical for evaluating this design.

Three major approaches appear in the literature:

ShaSpec (feature-level, CVPR 2023) is the most validated approach for shared-specific feature decomposition. The PDF provides comprehensive quantitative evidence: ShaSpec *achieves 86.26% average Dice across 15 modality combinations on BraTS2018 (Whole Tumor), with a gap of only 2.2 percentage points** from the full-modality baseline (88.46%). This near-parity with dedicated per-modality training is the most important quantitative finding — it demonstrates that shared-specific decomposition can achieve near-optimal performance with a single unified model. The DCO+DAO auxiliary loss framework is explicitly implemented in the HER2 prediction framework. The projection layer fθ_proj in the residual fusion (f = r + proj([s])) is what enables adaptation to arbitrary missing modality configurations — it learns to weight and combine available modality-specific features in a learned latent space. Ablation confirms that L1 loss is optimal for DAO (distribution alignment), and that auxiliary loss weights must be carefully tuned: α=0.1 (DCO) and β=0.02 (DAO) are optimal; larger values cause severe degradation.

DyMo (inference-time selection, 2026) offers a complementary approach: instead of generating missing features, it dynamically selects the best available modality combination at inference time. The information-theoretic motivation (using task loss as a proxy for modality relevance) is principled, but the approach has not been validated on histopathology WSI data.

MoRA (parameter-efficient fine-tuning, 2024) enables missing modality handling through efficient fine-tuning of pretrained models with <1.6% trainable parameters. The modality-aware up-projections are a clean design, but again, no histopathology validation.

Practical implication: The HER2 prediction framework's combination of CM-GAN (generation-level) + ShaSpec (feature-level) is more comprehensive than any single approach in the literature. The CM-GAN addresses the generation gap, while ShaSpec addresses the feature robustness gap. ShaSpec's 2.2pp cost for unified representation is an acceptable tradeoff given the flexibility gain.

Theme 3: Cross-Modal Synthesis — From GAN to Diffusion

Why this theme matters: The BCI dataset study establishes that Pyramid Pix2Pix (GAN-based) achieves PSNR=21.16dB, SSIM=0.477, but pathologists achieve only 37.5–40% accuracy on generated IHC images. This is the most critical clinical gap for the HER2 prediction framework's CM-GAN component.

The literature shows a clear transition from GAN-based to diffusion-based synthesis. Pyramid Pix2Pix (Liu et al., 2022) is the baseline with documented clinical acceptability limitations. HistDiST (arXiv 2025, PDF-confirmed) represents the current state of the art, achieving 28% MRA improvement on MIST Ki67 over MDCL GAN and 159% MRA improvement on BCI HER2 over MDCL. Its ablation provides the most granular understanding of what drives synthesis quality:

Ablation Component	FID	SSIM	Interpretation
SD+IC baseline	61.3	0.154	Poor quality without specialized components
+ v-prediction	37.5	0.198	v-prediction reduces FID by 39% — molecular fidelity improved dramatically
+ DDIM inversion	41.3	0.228	Inversion improves SSIM (structural fidelity) but raises FID (less realistic distribution)
+ Phikon cross-attention	31.3	0.227	Phikon embeddings fix FID degradation while maintaining SSIM gains

The HistDiST paper explicitly states it is "not clinically applicable at this stage." This establishes that even the current state of the art in stain transfer has not bridged the clinical validation gap.

Practical implication: Replacing CM-GAN (Pyramid Pix2Pix) with HistDiST-style diffusion synthesis could improve molecular fidelity, but would not solve the clinical validation gap. The HER2 prediction framework's approach of using synthesis to assist (not replace) pathologists is the right design principle given this evidence.

Theme 4: Domain Generalization — The Deployment Gap

Why this theme matters: All models in the HER2 prediction framework and the BCI dataset are trained and evaluated on data from a single institution. Cross-institutional deployment requires domain generalization.

AIDA (npj Precision Oncology 2024, PDF-confirmed) provides the most relevant evidence: +18.96 pp balanced accuracy improvement on cross-dataset breast cancer classification (ICIAR→BreaKHis). The FFT-Enhancer contributes +13.62 pp over adversarial training alone, demonstrating that frequency-domain processing is a powerful complement to pixel-space color normalization.

The Frontiers MRI paper's finding that frozen-weight feature extraction outperforms fine-tuning for cross-center generalization is directly relevant to the HER2 prediction framework's deployment strategy.

What does not transfer cleanly? AIDA is validated on binary benign/malignant classification, not HER2 molecular subtype prediction. The BCI dataset (breast

cancer, histopathology) is different from ICIAR/BreaKHis (also breast cancer, histopathology but different tissue types and staining protocols).

Practical implication: Domain adaptation is essential for clinical deployment, but the HER2 prediction framework has not yet incorporated it. The AIDA FFT-Enhancer should be evaluated as a preprocessing component.

Theme 5: Clinical Validation — The Most Critical Gap

Why this theme matters: The HER2 prediction framework's clinical value depends on pathologist acceptance of its outputs, not just algorithmic accuracy.

The BCI dataset study establishes the baseline clinical acceptability ceiling for synthesis-based approaches: pathologists achieve 37.5–40% accuracy on GAN-generated IHC images. This figure is cited here with full context — it represents the outcome of a study in which pathologists were asked to directly interpret synthetic IHC images generated from H&E input.

The Roche Navify study (Breast Cancer Research 2026) provides the most compelling clinical validation: F1 0.78→0.98 for HER2 ultra-low detection with AI assistance, including external validation on 89 slides from 2 hospitals. The key design principle — "AI assists, pathologist decides" — is explicitly validated as effective. Unlike direct image interpretation, AI-assisted scoring provides decision support without asking the pathologist to trust the generated images as standalone evidence.

Practical implication: The HER2 prediction framework's current design principle (AI assistance, not full automation) is the right approach given the clinical evidence. The next critical validation step should be a pathologist acceptance study using the framework's outputs, not just algorithmic accuracy benchmarks.

Theme 6: HER2-Low and Ultra-Low — The Clinically Most Consequential Gap

Why this theme matters: HER2-low and HER2-ultra-low cases (~10% of breast cancers) are clinically the most consequential because they qualify patients for ADC therapy. Most published HER2 AI studies focus on HER2+ vs. HER2- binary classification, not HER2-low specifically.

The Roche Navify study is the only study in this literature review specifically targeting HER2 ultra-low detection with external clinical validation. The dramatic precision improvement (0.39→0.91) demonstrates that AI can reduce the most clinically harmful error (underestimating HER2 expression).

Most other studies in this review are validated on binary or 4-way classification that includes HER2-low cases but do not specifically evaluate performance on this subgroup.

Practical implication: The HER2 prediction framework should include a HER2-low/ultra-low specific evaluation as part of its validation protocol.

5. Integrated Assessment for the Current Project

5.1 Most Justified Design Decisions

The HER2 prediction framework's core design decisions are well-supported by the literature:

1. **Multi-modal fusion (H&E+IHC)**: Convergent evidence from HER2 MAP (AUC 0.94), Frontiers MRI (AUC 0.94 with clinical data), and ViT Multistage (accuracy 0.94) confirms that multi-modal fusion substantially outperforms single-modality HER2 prediction.

2. **CM-GAN for missing IHC reconstruction**: The framework's CM-GAN addresses the most common clinical scenario (IHC unavailable). The baseline (Pyramid Pix2Pix, PSNR=21.16dB) is documented. HistDiST (diffusion) offers a potential upgrade path with 28% MRA improvement.

3. **ShaSpec-style shared-specific encoders**: The DCO+DAO auxiliary loss framework is explicitly implemented. The PDF confirms this as the most validated approach for feature-level missing modality handling across 15 modality combinations on BraTS2018.

4. **Channel Attention Module**: The within-framework ablation (95.32% vs. 92.41%) provides direct evidence. DyMo provides complementary support for adaptive feature weighting.

5. **"AI assists, pathologist decides" principle**: Validated by the Roche Navify study (F1 0.78→0.98, external validation on 89 slides, 2 hospitals).

5.2 Weakly Supported Assumptions

1. **Single-institution deployment**: No cross-institutional validation exists for the HER2 prediction framework. AIDA evidence (+18.96 pp on cross-dataset breast cancer) confirms this is a critical gap.

2. **HER2-low performance**: The framework has not been specifically evaluated on HER2-low or HER2-ultra-low cases, which are clinically the most consequential.

3. **ShaSpec on histopathology**: ShaSpec is validated on brain MRI (BraTS2018); its transfer to whole-slide histopathology is extrapolated from architectural principles, not empirical evidence.

5.3 Branches Deserving Further Testing

1. **Diffusion-based synthesis upgrade**: Replacing CM-GAN (GAN) with HistDiST (diffusion) could improve molecular fidelity. The 28% MRA improvement is measurable and would be expected to improve downstream HER2 scoring quality.

2. **AIDA FFT-Enhancer integration**: Adding frequency-domain domain adaptation preprocessing could improve cross-institutional robustness without architectural changes.

3. **CAM → DyMo integration**: Replacing the static CAM with DyMo's dynamic modality selection could improve adaptive fusion in ambiguous cases.

6. Unresolved Questions and Decision-Critical Gaps

Gap 1: Clinical Acceptability of CM-GAN Outputs

Why it matters: If the CM-GAN generates IHC images that pathologists would reject in clinical practice, the framework's missing-modality path may not be deployable.

Causal mechanism: The difficulty of achieving pathologist acceptance stems from a fundamental dissociation between pixel-level synthesis quality and diagnostic-relevant morphological features. Standard image quality metrics — PSNR, SSIM, FID — measure pixel-level fidelity, perceptual realism, and distribution-level similarity. However, pathologists rely on specific morphological features for HER2 scoring: membrane staining completeness, staining intensity patterns, and the spatial relationship between stained and unstained cells. These are highly structured, high-frequency features that may not correlate well with aggregate pixel-level metrics. GANs and diffusion models can produce visually plausible IHC textures — colors and textures that look like IHC to a human observer at a glance — while failing to preserve the membrane staining completeness patterns that pathologists use to distinguish HER2 0 from HER2 1+. This is why the BCI dataset study found pathologists achieving only 37.5–40% accuracy: the generated images look correct but lack the diagnostic-relevant morphological fidelity needed for scoring. Even HistDiST, which substantially improves molecular fidelity (159% MRA improvement), explicitly states it is not clinically applicable, confirming that improving pixel-level metrics does not automatically translate to pathologist acceptance.

What evidence is missing: No pathologist acceptance study has been conducted on the HER2 prediction framework's CM-GAN-generated IHC outputs specifically.

Practical feasibility of the validation study: A prospective CM-GAN pathologist acceptance study is a tractable near-term research problem, not merely a gap to be acknowledged. Three properties make it feasible:

1. **No new image acquisition required**: The same 977 test set pairs already used for algorithmic validation (BCI dataset, 51 patients, 4,870 patches, 3,896 train / 977 test split) can be presented to participating pathologists for independent scoring. The HER2 prediction framework's CM-GAN has already generated synthetic IHC for these cases; the remaining work is study design, IRB approval, and pathologist recruitment.

2. **A proven methodological template exists**: The Roche Navify study (Breast Cancer Research 2026) provides a concrete study design template: 188 cases, 9 participating pathologists, independent re-evaluation with and without AI assistance, expert consensus as ground truth, F1 improvement 0.78→0.98 for HER2 ultra-low detection, external validation on 89 slides from 2 hospitals. A CM-GAN acceptance study can follow this architecture — substituting the HER2 prediction framework's AI outputs for Navify's AI, and the BCI dataset test set for Navify's case cohort.

3. **Results directly answer Gap 1's core question**: A pathologist acceptance study with the BCI test set would produce exactly the evidence Gap 1 requires: whether pathologists, when presented with AI-assisted HER2 scores from the CM-GAN pipeline, achieve accuracy levels comparable to those achievable on real IHC images. This is the prerequisite determination before the more ambitious HistDiST multi-site clinical validation study can be justified.

Practical implication: Gap 1 is actionable in the near term. The study can be designed and executed within existing institutional resources, following an established template, and produces the single most decision-critical evidence item for the HER2 prediction framework's clinical deployment path.

Gap 2: Cross-Institutional Generalization

Why it matters: All evaluations use data from a single institution. Clinical deployment requires robustness to different staining protocols, scanners, and tissue processing procedures.

What evidence is missing: AIDA provides cross-dataset evidence for breast cancer classification (+18.96 pp), but not specifically for HER2 prediction or for the HER2 prediction framework's architecture.

What would reduce uncertainty: Evaluation on external datasets from institutions with different staining protocols. The AIDA FFT-Enhancer should be benchmarked as a preprocessing component.

Gap 3: HER2-Low and Ultra-Low Subgroup Performance

Why it matters: HER2-low and HER2-ultra-low cases are clinically the most consequential for ADC therapy eligibility. Underestimating HER2 expression in these cases has direct therapeutic consequences.

What evidence is missing: No specific evaluation on HER2-low/ultra-low subgroup accuracy. The Roche Navify study provides external validation for AI-assisted HER2 ultra-low detection, but not for the specific HER2 prediction framework.

What would reduce uncertainty: Subgroup analysis of HER2-low and HER2-ultra-low accuracy on the existing BCI dataset, with comparison to the Roche Navify study's benchmark (F1 0.98 for ultra-low detection).

Gap 4: End-to-End Joint Training of Synthesis + Classification

Why it matters: The current framework trains CM-GAN and the classifier separately. Joint training could allow the synthesis model to optimize for downstream classification quality rather than image quality metrics like PSNR.

Causal mechanisms: End-to-end joint training of synthesis and classification is technically difficult for four compounding reasons. First, **gradient misalignment** between generative and discriminative objectives: the synthesis model optimizes for reconstruction quality (minimizing pixel-level or perceptual loss between generated and real IHC), while the classifier optimizes for classification accuracy. These objectives can conflict — a synthesis model that produces visually perfect IHC images may not preserve the subtle morphological features most predictive of HER2 status, and conversely, the features most useful for classification may not correspond to visually realistic IHC synthesis. Second, **larger dataset requirements**: joint optimization of two deep networks requires substantially more training data than optimizing either alone, and the BCI dataset (4,870 patches) may be insufficient for stable joint training. Third, **training instability** arises from competing objectives — balancing the synthesis loss and the classification loss requires careful weighting that changes dynamically during training, and without proper scheduling, one objective can dominate and destabilize the other. Fourth, **metric misalignment**: the synthesis model over-optimizes for PSNR/SSIM/FID, which as discussed in Gap 1, are poorly correlated with diagnostic utility. A synthesis model trained jointly with classification could in principle learn to preserve diagnostically relevant features, but achieving this requires either a task-aware synthesis loss (e.g., using classifier gradients to guide synthesis) or a multi-task optimization framework that explicitly balances both objectives.

What evidence is missing: No literature on end-to-end joint training of synthesis + classification for HER2 prediction. ShaSpec (feature-level) and diffusion models (synthesis-level) are both trained separately from downstream classifiers.

What would reduce uncertainty: An ablation comparing separate training vs. joint training on the BCI dataset.

Gap 5: Generated IHC vs. Real IHC for Classification

Why it matters: The 95.09% dual-modality accuracy (real H&E+real IHC) vs. 94.25% H&E-only+CM-GAN accuracy represents a 0.84 percentage point gap. Is this gap clinically acceptable?

Clinical consequence magnitude: On the BCI dataset's test set (977 pairs), a 0.84pp accuracy gap translates to approximately **8 additional misclassifications** out of 977 test cases (977 × 0.84% ≈ 8.2). To assess clinical acceptability, this must be contextualized against two alternatives: (1) the baseline H&E-only accuracy of 71.44% represents approximately **279 misclassifications** out of 977 — the CM-GAN path reduces this by roughly 271

cases; (2) no IHC available at all means the patient either proceeds without HER2-targeted therapy (if HER2+) or receives unnecessary therapy (if HER2-). The 0.84pp gap between the CM-GAN path and real IHC is therefore small relative to the alternative of no synthesis at all. However, the clinical consequence of each misclassification is asymmetric: missing a HER2+ case (false negative) means denying targeted therapy; misclassifying HER2- as HER2+ (false positive) means exposing a patient to unnecessary therapy with associated toxicity. Without data on the direction of the 8 additional CM-GAN misclassifications, the clinical acceptability cannot be fully assessed. What can be said is that the CM-GAN path is substantially better than no synthesis (23.61pp improvement), and the 0.84pp gap between it and real IHC is likely clinically acceptable given the alternative.

What evidence is missing: No study has quantified the clinical consequence of this accuracy gap in terms of patient outcomes.

What would reduce uncertainty: A decision curve analysis comparing patient outcomes under HER2 prediction framework recommendations vs. standard clinical pathway, and a confusion matrix analysis to determine whether CM-GAN misclassifications are predominantly false positives or false negatives.

Gap 6: Foundation Model Integration

Why it matters: MuPD and CPath-Omni represent the most powerful and comprehensive multi-modal models for computational pathology. Could they replace or enhance the HER2 prediction framework's custom architecture?

What evidence is missing: Neither MuPD nor CPath-Omni has been evaluated on HER2-specific tasks. MuPD's 37% marker correlation improvement is promising but not validated on HER2. CPath-Omni's 15B parameters make it computationally prohibitive for routine use.

What would reduce uncertainty: Benchmarking CPath-Omni features or MuPD synthesis quality on the BCI dataset.

7. Recommended Next Steps

Step 1: Pathologist Acceptance Study for CM-GAN Outputs

Why it matters: The most critical clinical validation question is whether pathologists trust the framework's AI-assisted outputs. This is a prerequisite for clinical deployment.

Conduct a prospective study in which pathologists independently score HER2 status using (a) the HER2 prediction framework's outputs (with AI assistance) and (b) standard clinical workflow, with ground-truth HER2 status as reference. Compare accuracy, F1, and inter-observer agreement between conditions. Target: match or exceed the Roche Navify benchmark (F1 0.98 for ultra-low).

Step 2: Cross-Institutional Validation with AIDA FFT-Enhancer

Why it matters: Single-institution validation is insufficient for clinical deployment. The AIDA paper demonstrates that frequency-domain adaptation is effective for cross-institutional breast cancer classification.

Integrate the FFT-Enhancer as a preprocessing component in the HER2 prediction framework's pipeline. Evaluate on external datasets from institutions with different staining protocols. Target: maintain >94% accuracy on held-out institutional data.

Step 3: Diffusion Synthesis Upgrade (CM-GAN → HistDiST)

Why it matters: HistDiST achieves 28% MRA improvement and consistently lowest FID over GAN baselines. If molecular fidelity improves, downstream HER2 classification accuracy should also improve.

Replace the CM-GAN (Pyramid Pix2Pix) with a HistDiST-style LDM synthesis module. Evaluate on the BCI dataset. Compare downstream HER2 classification accuracy with and without diffusion synthesis. The HistDiST paper provides all architectural and training details needed for replication.

Step 4: HER2-Low and Ultra-Low Subgroup Analysis

Why it matters: HER2-low and HER2-ultra-low cases are the most clinically consequential subgroup. The Roche Navify study demonstrates that AI assistance dramatically improves ultra-low detection (F1 0.78→0.98).

Perform subgroup analysis of the HER2 prediction framework's existing BCI dataset results, stratifying by HER2 expression level (0, 1+, 2+, 3+). Quantify accuracy, F1, precision, and recall for each subgroup. Compare HER2-low performance against the Roche Navify external validation benchmark.

Step 5: CAM Ablation Expansion and DyMo Integration

Why it matters: The CAM ablation (95.32% vs. 92.41%) establishes that adaptive feature weighting matters. DyMo provides an information-theoretic alternative design for dynamic modality selection.

Expand the CAM ablation to include DyMo-style inference-time modality selection as an additional condition. Evaluate whether DyMo's principled information-theoretic approach outperforms CAM on ambiguous cases where both H&E and IHC are available but one is low-quality.

Step 6: End-to-End Joint Training Exploration

Why it matters: The current two-stage pipeline (synthesis then classification) may be suboptimal. Joint training could allow the synthesis model to optimize for classification outcomes rather than image quality metrics.

Design and implement an end-to-end training protocol that jointly optimizes the CM-GAN (or HistDiST-style diffusion) synthesis loss and the downstream HER2 classification loss. Compare against the current separate training protocol on the

BCI dataset. Key design consideration: incorporate task-aware synthesis losses (e.g., classifier gradient-guided synthesis) to bridge the gap between pixel-level quality and diagnostic utility.

8. Key Risks, Caveats, and Evidence Boundaries

Risk 1: Single-Institution Evaluation Scope

Risk: All performance numbers (95.09%, 94.25%, 90.28%) come from a single institution's dataset. This does not represent the diversity of clinical practice.

Evidence boundary: The AIDA paper demonstrates that domain shift causes substantial accuracy degradation (base 55.15% → 74.11% after adaptation) even for binary breast cancer classification. The HER2 prediction framework's cross-institutional accuracy is unvalidated.

Risk 2: Clinical Validation of Generated Images

Risk: The BCI dataset study establishes that pathologists achieve only 37.5–40% accuracy on GAN-generated IHC images. Even HistDiST (state of the art) explicitly states it is not clinically applicable.

Evidence boundary: No study has validated pathologist acceptance of the HER2 prediction framework's specific CM-GAN outputs. The clinical acceptability ceiling for synthesis-based HER2 assessment is not established.

Risk 3: HER2-Low and Ultra-Low Representation

Risk: The BCI dataset may under-represent HER2-low and HER2-ultra-low cases, which are the clinically most consequential subgroups for ADC therapy eligibility.

Evidence boundary: No specific HER2-low or HER2-ultra-low subgroup performance data is available from the grounded materials. While the BCI dataset contains 4,870 patches across 51 patients, the distribution of HER2 expression categories within this cohort is not reported in the source documents. The Roche Navify study (P004, Breast Cancer Research 2026) provides external validation for this subgroup with 188 cases, but uses a different AI system (Roche Navify) and a different patient cohort — the HER2-low case prevalence in the Navify cohort (derived from ~10% HER2-low rate in general breast cancer population) cannot be directly mapped to the BCI dataset's composition. Specific HER2 expression level distribution data for the BCI dataset would be needed to assess whether this risk is significant or minor.

Risk 4: Small Dataset for ShaSpec Pathology Validation

Risk: ShaSpec is validated on BraTS2018 (brain MRI) and Audiovision-MNIST. Its performance on histopathology whole-slide images (different modality, different scale, different tissue characteristics) is untested.

Evidence boundary: The HER2 prediction framework implements ShaSpec's design, but no direct histopathology validation of ShaSpec exists in the literature. The assumption that it transfers is extrapolated from the architectural principles.

Risk 5: Commercial AI Competition

Risk: The Roche Navify system (commercial, not openly available) has demonstrated strong external validation for HER2 ultra-low detection. The HER2 prediction framework, if deployed, would compete with an established commercial product.

Evidence boundary: Head-to-head comparison between the HER2 prediction framework and Roche Navify has not been conducted. The Roche Navify study (P004) establishes that AI-assisted HER2 scoring achieves F1 0.78→0.98 for ultra-low detection with external validation — the most rigorous clinical validation benchmark in this review. Beyond the lack of head-to-head comparison, three specific competitive barriers exist: (1) regulatory — Roche Navify has received CE/FDA regulatory clearance pathways; academic systems require independent regulatory evaluation; (2) integration — Roche Navify is deployed within the Roche Navify Digital Pathology platform, enabling seamless clinical workflow integration; (3) trust — commercial products with published clinical validation have established clinician trust that academic systems must rebuild. The HER2 prediction framework's competitive position would benefit from direct benchmarking against Navify performance metrics and explicit regulatory pathway planning.

Risk 6: Overclaiming Synthesis Quality

Risk: The CM-GAN achieves 94.25% H&E-only accuracy, which may create an impression of clinical reliability. However, the BCI dataset study established that pathologists achieve only 37.5–40% accuracy on generated images — meaning the generated IHC is not independently interpretable.

Evidence boundary: The 94.25% figure is an algorithmic classification accuracy on a test set, not a measure of clinical trustworthiness. These should not be conflated in communications about the framework.