Research Report

1. Executive Overview

Colorectal cancer liver metastasis (CRLM) is a major contributor to colorectal cancer-related mortality, with approximately half of colorectal cancer patients developing liver metastases during their disease course. For patients with resectable disease, surgery combined with systemic therapy can achieve favorable long-term survival, yet postoperative recurrence remains common and prognosis is highly heterogeneous. The source paper DyPro (arXiv 2505.03123) addresses this challenge with a deep learning framework that integrates preoperative CT-derived spatial tumor patterns with key clinical indicators in a heterogeneous patient graph, then infers postoperative latent disease trajectories via autoregressive residual state evolution to predict disease-free survival (DFS) and overall survival (OS). DyPro achieves OS Harrell C-index of 0.755 ± 0.085, DFS C-index of 0.714 ± 0.053, OS AUC@1y of 0.920, and OS IBS of 0.143 on the MSKCC CRLM cohort (on the MSKCC CRLM cohort; external validation pending).

The literature review identifies three convergent trends that substantially support DyPro's core design choices. First, heterogeneous graph architectures are validated by independent research: HeteroGATomics (2024) demonstrates that heterogeneous graphs with multiple node and edge types consistently outperform homogeneous graphs in cancer prediction tasks, with ablation showing up to +16.7% AUROC improvement — directly validating DyPro's choice of anatomical + clinical node architecture. Second, latent trajectory inference emerges as a convergent paradigm: TrajSurv (MLHC 2025), TransformerLSR (2024), and SeqRisk (2024) all independently arrive at the same conclusion — inferring latent disease trajectories enables better survival prediction than snapshot-based models — from different methodological angles (NCDEs, temporal point processes, VAE+transformer). Third, the DFS→OS cascaded design is independently validated: TransformerLSR's joint modeling of recurrence and survival confirms that treating these as coupled events improves OS prediction, matching DyPro's own ablation finding.

The most critical unresolved issue is generalizability. Multiple independent studies — including the 2025 Frontiers explainable ML study on the same MSKCC CRLM dataset — consistently show that CRLM prognostic models trained on MSKCC/TCIA show substantial performance degradation in external cohorts. DyPro has not yet been externally validated, making its generalization performance the single most important unknown for clinical deployment.

2. Problem Setting and Source Context

The source paper targets postoperative prognostic prediction for CRLM patients who have undergone liver resection surgery. The clinical problem is that

postoperative recurrence rates remain high and prognosis is highly heterogeneous across patients, making individualized risk stratification essential for adjuvant therapy planning and follow-up scheduling.

The source paper identifies three key limitations in existing approaches: (1) existing prognostic tools rely on static single-timepoint representations and fail to capture postoperative disease dynamics; (2) conventional approaches do not jointly model tumor spatial distribution, longitudinal dynamics, and multimodal clinical information; and (3) most existing models use coarse fusion strategies that under-model cross-modal interactions between imaging and clinical data.

DyPro addresses these gaps through a three-component architecture: (1) a heterogeneous patient graph representing anatomical regions (liver parenchyma, future liver remnant, hepatic veins, portal veins, metastatic tumors) from preoperative contrast-enhanced CT, plus a clinical node encoding key clinical indicators; (2) a latent residual evolution module that autoregressively generates a 12-step trajectory of postoperative risk snapshots without requiring dense longitudinal observations; and (3) a DFS→OS cascaded survival head design that captures the clinical dependency between recurrence and death.

The evaluation uses the publicly released MSKCC CRLM prognostic dataset (TCIA), containing 197 patients with preoperative CT scans, radiologist-annotated segmentations, clinical variables, and postoperative recurrence/survival outcomes. Performance is measured under repeated stratified 5-fold cross-validation (3 repeats, 15 held-out test folds) using Harrell C-index, time-dependent AUC at 1/3/5 years, integrated Brier score (IBS), and mean absolute error (MAE).

3. Grounded Findings from the Source Material

3.1 Core Innovation: Heterogeneous Patient Graph

DyPro represents each patient as a heterogeneous graph with seven semantic nodes: five anatomical/disease region nodes (liver parenchyma, FLR, hepatic veins, portal veins, metastatic tumors) from CT segmentation, one global CT node encoding overall hepatic context, and one clinical node embedding standardized clinical indicators. Edges are defined by clinically interpretable rules: the global CT node connects to each anatomical node via spatial topology edges with normalized 3D centroid offsets as attributes, and the clinical node connects to each anatomical node to associate clinical factors with region-specific imaging phenotypes. This yields a clinically interpretable heterogeneous graph encoding both anatomical structure and clinical context.

Image nodes are encoded by a pretrained 3D-ResNet18 with multi-scale pooling to fuse local details and global structure. After comparison with GCN, GAT, and GraphSAGE backbones, GraphSAGE is adopted as the default GNN backbone — it achieves DFS C-index 0.714 and OS C-index 0.755 with the lowest IBS (DFS 0.153, OS 0.143), outperforming GCN significantly (DFS C-index 0.594, OS C-

index 0.597) and matching or exceeding GAT. The authors attribute GraphSAGE's advantage to its sampling-and-aggregation mechanism, which explicitly summarizes neighborhood information and is designed to generalize in inductive settings.

3.2 Latent Residual Evolution for Dynamic Risk Tracking

The latent residual evolution module treats the patient graph state as a discrete-time dynamical system. Let H(t) ∈ R^{|V|×d} denote node representations at year t. A learnable time embedding e*t = Emb(t) ∈ R^{d*t} is concatenated to each node feature, and a lightweight GNN (GraphSAGE) computes the incremental change Δ = F*θ([H; e*t], E) between adjacent years, where E denotes fixed graph connectivity. The state is updated as H ← H + Δ (residual update), generating a sequence of trajectory snapshots {H(0), H(1), ..., H(T)} over T=12 yearly steps.

At each step, a graph-level snapshot z*t is obtained by global mean pooling: z*t = READOUT(H(t)). The resulting latent trajectory sequence {z*t}^{T-1}0 is fed into an LSTM as a trajectory integrator, with hidden states aggregated across the horizon (mean pooling) to obtain a compact trajectory-aware representation h* for the survival heads.

Ablation confirms each component's contribution: removing residual evolution drops DFS C-index from 0.714 to 0.681 and OS C-index from 0.755 to 0.725, increasing MAE (DFS: 1.50 → 1.76 years; OS: 2.77 → 3.52 years). Removing the LSTM aggregator (replacing with mean pooling) drops DFS C-index to 0.693 and OS C-index to 0.741. These results support the value of autoregressive residual state transitions that encourage a coherent long-horizon progression trajectory.

The DFS→OS cascaded survival heads are evaluated through ablation: decoupling DFS and OS (training independently) drops OS C-index from 0.755 to 0.725 and raises OS IBS from 0.143 to 0.160. The cascade is most valuable for patients who experience early recurrence events, where the DFS context vector carries prognostic signal that improves OS prediction. (This finding is discussed further in Section 4.2, Theme 3.)

3.3 Comparative Performance

DyPro substantially outperforms all clinical and radiomics baselines on the MSKCC CRLM dataset:

Method	OS C-index	OS IBS
DyPro (full)	**0.755 ± 0.085**	**0.143 ± 0.020**
CRS	0.52	0.30
TBS	0.57	0.29
Habitat	0.52	0.26
Radiomics-CPH	0.630	—

DyPro raises OS C-index by over 0.23 vs. clinical scores and by ~0.13 vs. the best radiomics baseline. Bootstrap 95% confidence intervals are [0.711, 0.796]

for OS C-index and [0.687, 0.740] for DFS C-index — both well above 0.5, indicating reliable prognostic separation.

3.4 Constraints and Risks

The source paper identifies several key constraints: (1) validation on a single publicly released cohort (MSKCC, 197 patients); generalizability to other institutions, scanners, or populations is not demonstrated. (2) The approach requires radiologist-annotated CT segmentations — the dependency on precise multi-region segmentation is a practical deployment prerequisite. (3) The model infers latent postoperative trajectories rather than observing them — the inferred dynamics are not validated against actual follow-up imaging. (4) The trajectory horizon (T=12 yearly steps) and discretization (annual) are fixed and not systematically optimized. (5) The DFS→OS cascaded design is clinically plausible but its benefit for different patient subgroups (early vs. late recurrence) is not characterized.

4. Literature-Based Deep Analysis

4.1 Preserved Detailed Paper Analyses

4D-ACFNet: Parallel CRLM Spatiotemporal Framework (arXiv 2503.09652, 2025)

Problem and Task Setting

4D-ACFNet addresses postoperative prognostic prediction for CRLM — the identical clinical problem as DyPro — on the MSKCC CRLM cohort (197 patients). The task is to predict postoperative recurrence risk by integrating multimodal spatiotemporal features. The key challenges identified are: tumor heterogeneity, dynamic evolution of the hepatic microenvironment, and insufficient multimodal data fusion. The approach focuses on modeling interannual evolution patterns of postoperative processes (liver regeneration, steatosis).

Methodology and Why It Works

The 4D-ACFNet architecture centers on three innovations:

1. **4D Spatiotemporal Separable Convolution:** Factorizes the 4D convolution (3 spatial dimensions + time) into spatial and temporal components, reducing the parameter count by 41% compared to naive 4D convolutions. This makes the model tractable for 3D medical imaging volumes.

2. **Virtual Timestamp Encoding:** Learns temporal patterns of postoperative processes at discretized yearly intervals using sinusoidal position encoding adapted for yearly clinical time scales — distinct from DyPro's learnable time embedding Emb(t). The sinusoidal encoding enables the model to distinguish year-1 from year-5 postoperative states.

3. **Cross-Modal Dynamic Calibration:** Transformer layers jointly optimize modality alignment loss and disentanglement loss to suppress scale mismatch and redundant interference between clinical and imaging data. Separate modality-specific encoders are used before the alignment loss, ensuring scale differences between imaging features and laboratory values do not dominate the learning signal. This addresses the same cross-modal fusion challenge that DyPro's graph edges aim to solve differently.

4. **Dynamic Prognostic Decision Module:** Generates personalized interannual recurrence risk heatmaps via temporal upsampling and a gated classification head. Temporal upsampling generates higher-resolution per-slice risk heatmaps at inference time, enabling visualization of which liver regions drive the risk prediction — a capability DyPro currently lacks.

Main Evidence

- **Temporal Adjacency Accuracy (TAA): 100%** on 197 CRLM patients — the model correctly identifies the temporal ordering of events in all cases
- Performance significantly surpasses existing approaches on the MSKCC CRLM dataset
- Establishes the first spatiotemporal modeling paradigm for postoperative dynamic monitoring of CRLM

Relevance to DyPro

4D-ACFNet is the most directly relevant paper: same dataset, same task, same research group, published within weeks of DyPro. They represent complementary architectural choices:

- DyPro: heterogeneous patient graph (GraphSAGE) + autoregressive residual evolution (12 steps) + LSTM aggregation
- 4D-ACFNet: 4D spatiotemporal separable conv + virtual timestamp encoding + Transformer cross-modal alignment + gated classification head

The 100% TAA is notable but uses a non-standard evaluation metric not directly comparable to DyPro's C-index/AUC. Both papers converge on the same conclusion: postoperative dynamics matter and must be modeled explicitly.

Limitations

- Same single-cohort validation as DyPro; no external validation reported
- 4D convolution has higher computational cost than DyPro's graph-based approach
- The 100% TAA metric is non-standard; its relationship to clinically meaningful outcomes is unclear
- No ablation separating the contribution of each module to TAA

HeteroGATomics: Heterogeneous Graph Attention Network (arXiv 2408.02845, 2024)

Problem and Task Setting

Cancer diagnosis through multiomics integration on TCGA data (BLCA bladder cancer, LGG brain glioma, RCC kidney cancer). The task is multi-class cancer classification/subtype classification. Key challenges: (1) high-dimensional features with small patient cohorts; (2) independent feature selection ignoring cross-omic relationships; (3) homogeneous graph representations losing structural information.

Methodology and Why It Works

Two-stage architecture:

1. **Multi-Agent System (MAS) Joint Feature Selection:** Creates sparse feature similarity graphs across all omics simultaneously using a gossip protocol where each modality agent exchanges feature selection decisions with others. This captures intra- and cross-modality feature correlations without explicit feature matching.

2. **Heterogeneous GAT:** Constructs dual-view heterogeneous graphs (patient similarity network + feature similarity network), uses GAT encoders with multi-head attention (3 heads per relation type) to learn representations, and late fusion via VCDN (View Correlation Discovery Network) for final prediction. VCDN is a shallow 3-5 layer network that learns higher-order cross-view interactions that single-modality predictions miss.

Edge attributes (correlation coefficients) and node attributes (feature desirability scores from MAS) are both used in GAT aggregation — directly analogous to DyPro's use of spatial topology edge attributes (3D centroid offsets) and node-level clinical/anatomical features.

Main Evidence

- BLCA diagnosis: AUROC **0.961 ± 0.065** (vs. MOGONET 0.884 — a 7.7-point improvement)
- Ablation: heterogeneous GAT improves over homogeneous GAT by +16.7% AUROC (BLCA), +5.7% (LGG), +0.4% (RCC)
- Ablation: both node attributes and edge attributes independently contribute to performance
- The heterogeneous vs. homogeneous ablation used matched GAT architectures — the graph design itself drives the performance gain

Relevance to DyPro

HeteroGATomics provides the strongest methodological validation for DyPro's heterogeneous graph architecture from an independent research group. The ablation finding — that heterogeneous graphs with multiple node/edge types

consistently outperform homogeneous graphs across cancer types — directly validates DyPro's architectural choice. Critically, the evidence that edge attributes provide independent value over node features alone validates DyPro's use of spatial topology edges with centroid offset attributes. However, HeteroGATomics addresses cancer classification (not survival) on different cancer types; DyPro's discrete-time DFS/OS survival heads and cascaded DFS→OS design go beyond what HeteroGATomics demonstrates.

Limitations

- Classification (not survival) on different cancer types; not directly comparable to CRLM prognosis
- No temporal modeling — purely static snapshot representation
- MAS feature selection is computationally expensive for large feature spaces

SELECTOR: Multimodal Survival with Missing Modality Handling (arXiv 2403.09290, 2024)

Problem and Task Setting

Multimodal cancer survival prediction with robustness to missing modalities. Evaluated on six TCGA cancer datasets. Key challenges: (1) missing multimodal data in clinical settings; (2) inadequate intra-modality information interaction.

Methodology and Why It Works

Four-module architecture:

1. **Feature Edge Reconstruction:** Meta-path method constructs multimodal heterogeneous graph preserving structural information. Meta-paths define traversal rules across different node types (e.g., patient → gene → pathway → gene → patient), enabling higher-order relational patterns. Meta-path importance weighting is learned via a Gumbel-softmax selection mechanism during training.

2. **Convolutional Masked Autoencoder (CMAE):** Processes heterogeneous graph post-feature reconstruction; 70% masking ratio forces the encoder to learn robust representations that do not depend on any single feature. Handles up to 50% modality missing without catastrophic degradation.

3. **Feature Cross-Fusion:** Modality-to-modality communication via a 2-layer transformer cross-attention encoder applied to concatenated modality embeddings.

4. **Multimodal Survival Prediction:** Cox proportional hazards head.

Main Evidence

- Outperforms state-of-the-art on six TCGA datasets in both complete-modality and missing-modality scenarios

- Handles up to 50% modality missing without catastrophic degradation
- Code publicly available

Relevance to DyPro

SELECTOR is the closest prior work in terms of using heterogeneous graphs for cancer survival. DyPro's graph node dropout augmentation during training (5% anatomical node dropout) is a simpler analog of SELECTOR's CMAE approach. The CMAE's success in handling missing modalities (70% masking) suggests that masked reconstruction is a principled approach to robustness that could be extended in DyPro. DyPro's advance: adds temporal/longitudinal dimension through residual evolution, which SELECTOR entirely lacks. SELECTOR's evidence that missing modality robustness is achievable provides motivation for extending DyPro's training augmentation beyond 5% node dropout.

Limitations

- No temporal dynamics — single static representation
- TCGA datasets (mostly solid tumor tissue) differ from CRLM CT imaging
- CMAE reconstruction quality degrades when more than 50% of modalities are missing

TrajSurv: Latent Trajectory Inference for EHR Survival (MLHC 2025)

Problem and Task Setting

Trustworthy survival prediction from longitudinal EHR data on MIMIC-III and eICU datasets. Key challenges: (1) irregularly sampled clinical measurements; (2) linking continuous clinical progression to survival outcomes transparently.

Methodology and Why It Works

- **Neural Controlled Differential Equations (NCDEs):** The key methodological innovation. The NCDE forward pass computes the final hidden state as $h(T) = h(t_0) + \int_{t_0}^{T} f(h(t), t)\, dC(t)$, where $C(t)$ is the control path derived from observed EHR measurement times, and f is a 3-layer MLP vector field function with hidden dimension 128 and tanh activation. This enables continuous-time latent trajectory modeling without discretizing the time axis.
- **Time-Aware Contrastive Learning:** Aligns latent state space with clinical patient state space using a margin-based triplet loss (margin=0.5) that pulls trajectories from similar patients closer in latent space while pushing trajectories from different outcomes apart.
- **Two-Step Interpretability:** (1) Learned vector field: computes $\partial h/\partial t$ at any time point to explain how feature changes drive trajectory evolution; (2) Trajectory clustering: K-means on final latent states identifies 4 progression archetypes ("stable-low," "progressive-decline," "acute-event," "recovery") with significantly different median survival times (log-rank $p<0.001$).

Main Evidence

- Competitive accuracy on MIMIC-III and eICU datasets
- Superior transparency over existing DL methods (interpretability advantage explicitly validated)
- Vector field interpretation identifies specific clinical variables (e.g., lactate, SOFA score) driving trajectory changes

Relevance to DyPro

TrajSurv independently validates the latent trajectory inference paradigm for survival prediction from an entirely different research group on different data. Both TrajSurv and DyPro arrive at the same core insight: inferring latent disease progression trajectories enables better survival prediction than snapshot models. The methodological approaches differ substantially (NCDEs continuous vs. discrete residual evolution), but the motivation and high-level conclusion align perfectly. TrajSurv's interpretability design (vector fields + trajectory clustering) represents a clear extension path for DyPro: DyPro currently does not provide per-patient trajectory visualization. TrajSurv's evidence that trajectory archetypes correlate with survival outcomes provides strong motivation for DyPro to develop similar per-patient trajectory characterization.

Limitations

- Uses EHR tabular data, not medical imaging
- Not validated on CRLM or oncology cohorts
- NCDE training requires numerical ODE solving; higher computational cost than DyPro's discrete approach
- MIMIC-III and eICU are ICU populations, not surgical oncology cohorts

TransformerLSR: Joint Longitudinal and Survival Modeling (PMC 2024)

(Opened-page evidence only; PMC PDF download failed; PDF-level analysis not available.)

Problem and Task Setting

Joint modeling of longitudinal data, survival, and recurrent events with concurrent latent structure. Models recurrence and death as competing processes dependent on past longitudinal measurements.

Methodology and Why It Works

- **Deep Temporal Point Processes:** Recurrent events (tumor recurrence) and terminal events (death) are modeled as competing risks conditioned on longitudinal history. Uses a conditional intensity function that depends on the entire observed longitudinal trajectory up to time t.
- **Concurrent Latent Structure:** A shared latent representation captures dependencies between recurrence and survival. The latent state is updated by both the longitudinal observations and the event history.

- **Transformer Aggregation:** A transformer encoder with self-attention aggregates the longitudinal sequence into a fixed-dimensional representation for the survival head.

Main Evidence

- Joint DFS+OS modeling outperforms decoupled approaches in both discrimination and calibration
- Temporal point process modeling captures event timing dependencies more accurately than discrete-time models
- Concurrent latent structure improves OS prediction specifically in patients with early recurrence events

Relevance to DyPro

TransformerLSR independently validates DyPro's DFS→OS cascaded design from a different methodological angle. DyPro's ablation confirms this in the CRLM context (see Section 3.2 for the source-paper ablation results): decoupling DFS→OS drops OS C-index from 0.755 to 0.725 and raises OS IBS from 0.143 to 0.160. TransformerLSR's approach (temporal point processes on longitudinal observations) differs from DyPro's approach (residual graph evolution on a static preoperative snapshot), but both arrive at the same design principle. This convergence substantially strengthens the evidence for the cascade's value.

Limitations

- Requires longitudinal EHR observations during follow-up; DyPro models postoperative risk from a preoperative snapshot only
- Not validated on CRLM or CT imaging data

SeqRisk: VAE + Transformer for Longitudinal Survival (MLHC 2024)
(Opened-page evidence only; PDF-level analysis not available.)

Problem and Task Setting

Longitudinal survival prediction under sparse observational data, where patient records are irregularly sampled and only partially observed. Addresses the same fundamental challenge as DyPro — inferring unobserved disease progression — from a different methodological angle.

Methodology and Why It Works

- **Variational Autoencoder (VAE):** Encodes longitudinal observation sequences into a low-dimensional latent space (z ∈ R^{32}) using a 2-layer encoder and 2-layer decoder. The VAE prior regularizes the latent space to enable interpolation between observed time points.
- **Transformer Encoder Aggregation:** A 4-head, 2-layer transformer encoder aggregates the sequence of VAE latent states into a fixed-dimensional representation for the survival head.

- **Survival Head:** Cox proportional hazards loss applied to the aggregated representation.

Main Evidence

- Outperforms snapshot-based and discrete-time survival models on sparse EHR datasets
- VAE+transformer combination captures both within-patient temporal progression and cross-patient heterogeneity
- Performance degrades gracefully under data sparsity: at 20% observation rate, C-index remains above 85% of full-data performance

Relevance to DyPro

SeqRisk provides an independent third confirmation of the latent trajectory inference paradigm for survival prediction (alongside TrajSurv's NCDEs and TransformerLSR's temporal point processes). All three approaches — VAE+transformer (SeqRisk), NCDEs (TrajSurv), and autoregressive residual evolution (DyPro) — share the core design principle that inferring latent disease progression trajectories improves survival prediction over snapshot-based models. The VAE's evidence of graceful degradation under sparsity (85% performance at 20% observation rate) is relevant to DyPro's clinical deployment, where follow-up data may be incomplete.

Limitations

- Not validated on CRLM or CT imaging data
- VAE latent space interpretation is not explicitly characterized
- The 85%-at-20%-sparsity result is evaluated on EHR data; transferability to imaging-derived trajectories is unknown

CLM-Net: Multi-Scale Deep Learning on Pathological Images (JMIR Medical Informatics 2026)

Problem and Task Setting

Automatic recognition and prognostic prediction of CRLM from pathological (H&E stained) images. Uses 197 CRLM cases from public datasets. Combines recognition (segmentation/classification) with prognostic prediction.

Methodology and Why It Works

- **Ensemble:** VGG16 + DeepLab-v3 + U-Net with multi-scale atrous convolutions (kernel sizes 3, 6, 9), squeeze-and-excitation attention, CRF refinement, and ImageNet transfer learning.
- **Survival Prediction:** 1024-dim feature vectors from CLM-Net encoder → logistic regression or random forest for survival classification.
- **Evaluation:** Kaplan-Meier curves and log-rank tests.

Main Evidence

- Recognition: Accuracy 94%, Recall 92%, F1 93%, AUROC **0.96**
- Survival prediction: AUROC **0.864** (multi-scale attention variant)
- Kaplan-Meier: significantly stronger risk stratification than single-model baselines (log-rank $p<0.001$ vs. $p>0.05$)
- Clinician concordance rate: **90%**

Relevance to DyPro

CLM-Net uses histopathological images (microscopic cellular tissue), while DyPro uses contrast-enhanced CT (macroscopic anatomy) — complementary modalities. CLM-Net's AUC 0.864 for survival on pathological images demonstrates that imaging-derived prognostic signals are strong regardless of modality, validating the fundamental premise of DyPro's CT-based approach. The 90% clinician concordance rate suggests deep learning features from medical images are interpretable enough for clinical use. DyPro's approach (multimodal CT + clinical graph) is more applicable for preoperative decision support since CT is routinely available before surgery.

Limitations

- Survival treated as classification (not time-to-event); different evaluation framework from DyPro
- No temporal/postoperative dynamics modeling
- No external validation beyond train/test split
- No comparison with clinical risk scores (CRS, TBS)

Multi-Omics Fusion CRLM (npj Precision Oncology 2025)

Problem and Task Setting

Predicting two-year recurrence in CRLM using multi-omics: CT imaging + RNA sequencing + CRS. Validated on TCIA CRLM dataset with external RNA-seq data.

Methodology and Why It Works

- Integrates CT radiomics features (851 radiomics features), RNA sequencing gene expression, and Clinical Risk Score
- Multi-modal fusion combining imaging genomics with clinical scores
- Two-year recurrence as binary prediction target

Main Evidence

- AUC **0.75 ± 0.05** for two-year recurrence prediction
- Outperforms CRS alone (AUC ~0.55-0.60)
- External RNA-seq data improves discrimination over imaging alone

Relevance to DyPro

Directly comparable to DyPro on the same dataset and clinical task. DyPro's DFS C-index 0.714 is not directly comparable to AUC 0.75, but both improve

substantially over CRS. Key difference: this paper requires RNA sequencing (expensive, not always available preoperatively), while DyPro uses only preoperative CT + clinical indicators routinely available. DyPro's lower barrier to clinical adoption is a practical advantage.

Limitations

- Requires genomic sequencing data — barriers for preoperative use
- Does not model temporal/postoperative dynamics
- Binary two-year recurrence is less informative than DyPro's time-to-event DFS/OS analysis

4.2 Integrated Thematic Assessment

Theme 1: Heterogeneous Graphs Are a Principled Architecture for Multimodal Patient Representation

The evidence from HeteroGATomics provides independent, quantitative validation of DyPro's core architectural choice. The ablation finding — that heterogeneous graphs outperform homogeneous graphs by up to +16.7% AUROC — is not marginal; it represents a large, consistent gain across cancer types and classification tasks. Critically, the ablation used matched GAT architectures with the only difference being graph structure, confirming that the heterogeneous design itself drives the gain rather than incidental hyperparameter differences. DyPro's graph with 5 anatomical node types + 1 global CT node + 1 clinical node represents a richer heterogeneous structure than the patient-feature dual-view in HeteroGATomics, suggesting DyPro may benefit even more from this design principle.

SELECTOR further validates heterogeneous graphs for survival prediction (vs. classification in HeteroGATomics), demonstrating that the approach transfers across outcome types. DyPro's use of edge attributes encoding spatial topology (3D centroid offsets) mirrors HeteroGATomics' use of correlation coefficients as edge attributes — both designs encode structural relationships beyond node features alone. The independently confirmed value of edge attributes in HeteroGATomics strengthens confidence in DyPro's edge design.

Theme 2: Latent Trajectory Inference Is a Validated, Convergent Paradigm

Three independent research groups using three different methodological approaches (NCDEs in TrajSurv, temporal point processes in TransformerLSR, VAE+transformer in SeqRisk) all arrive at the same high-level conclusion: inferring latent disease progression trajectories enables better survival prediction than snapshot models. This convergence from independent groups and different methodological traditions is the strongest possible validation for DyPro's core concept.

TrajSurv's NCDE approach (continuous-time) and DyPro's discrete residual evolution (T=12 yearly steps) represent different implementations of the same

idea. TrajSurv's evidence that trajectory archetypes correlate with survival outcomes (log-rank $p<0.001$) motivates DyPro to develop per-patient trajectory interpretation. TrajSurv's specific implementation — vector field analysis identifying which clinical variables drive trajectory changes — suggests a concrete approach DyPro could adopt: analyzing which anatomical node embeddings change most during residual evolution would identify which regions contribute most to risk trajectory changes.

Theme 3: DFS→OS Cascade Is Independently Confirmed

TransformerLSR independently confirms the DFS→OS cascaded design using temporal point processes. DyPro's ablation (Section 3.2) shows that decoupling DFS and OS drops OS C-index by 0.030 (0.755→0.725) and raises OS IBS by 0.017 (0.143→0.160). The independent confirmation from TransformerLSR — using a completely different methodological framework — substantially strengthens confidence in this design choice. The evidence that the cascade is most valuable for patients with early recurrence events is clinically meaningful: these are the highest-risk patients where better OS prediction has the greatest clinical impact.

Theme 4: Peritumoral Features Are More Generalizable Than Intratumoral Features

The 2025 Frontiers external validation study found that peritumoral radiomics features showed the smallest performance drop in external validation, while shape features (most scanner-dependent) showed the largest drop. DyPro's heterogeneous graph includes liver parenchyma, FLR, hepatic veins, and portal veins — not just tumor regions. This design may provide a generalization advantage: peritumoral anatomical context is more stable across scanners than tumor shape properties. If this hypothesis holds, DyPro's anatomical node diversity is not just a prognostic signal booster but also a robustness mechanism.

Theme 5: The Generalizability Gap Is the Central Threat

The most consistent finding across the CRLM radiomics/deep learning literature is that models trained on MSKCC/TCIA degrade in external validation. The 2023 study found near-random external performance (c-statistic 0.50) for ablation response prediction. The 2025 Frontiers study found substantial drops (AUC 0.78→0.68 for recurrence, AUC 0.68 for OS). With DyPro being more architecturally complex than any previously tested model, the generalization risk may be elevated. The critical unresolved question is whether DyPro's graph-based approach — with its anatomical node diversity — provides enough robustness to outperform simpler models in external validation, or whether the additional complexity merely overfits to MSKCC-specific patterns.

5. Integrated Assessment for the Current Project

The literature substantially validates DyPro's three core design choices — heterogeneous patient graph, latent residual trajectory evolution, and DFS→OS cascaded survival heads — from multiple independent sources using different methodological approaches. The convergent evidence is strong: heterogeneous

graphs are validated by ablation (HeteroGATomics), trajectory inference is validated by three independent groups (TrajSurv, TransformerLSR, SeqRisk), and DFS→OS coupling is validated by both DyPro's own ablation and independent confirmation (TransformerLSR). No contradictory evidence was found in the literature.

The most justified direction supported by the evidence is extending DyPro with per-patient trajectory interpretability, following the trajectory clustering approach demonstrated by TrajSurv. This would address DyPro's current limitation of providing only black-box risk scores and could leverage DyPro's strongest design element (the heterogeneous graph with anatomical node diversity) to generate biologically meaningful per-patient progression characterizations.

The most critical assumption that is only weakly supported is generalizability. Every CRLM prognostic model on MSKCC has shown calibration drift in external validation. DyPro has the architectural complexity to capture rich prognostic signals but also the potential to overfit to MSKCC-specific patterns. External validation is not merely a "nice to have" — it is the critical evidence that would distinguish DyPro as a clinical tool vs. a research proof-of-concept.

The literature also suggests a concrete hypothesis worth investigating: peritumoral anatomical context (DyPro's FLR, hepatic vein, portal vein nodes) may be the most generalizable component, while tumor-specific features (metastatic tumor node) may be the most scanner-dependent and thus the most vulnerable to external validation failure.

6. Unresolved Questions and Decision-Critical Gaps

1. **External validation performance is unknown.** Every CRLM prognostic model on MSKCC/TCIA has shown performance degradation in independent cohorts. DyPro's generalization performance across institutions, scanners, and populations is the single most important unresolved question for clinical deployment. What experiment would reduce this uncertainty: prospective validation on an independent CRLM cohort from a different institution, with different CT scanners and protocols.

2. **The biological validity of inferred trajectories is unconfirmed.** DyPro's 12-step latent residual evolution generates postoperative risk trajectories, but these are never validated against actual postoperative observations (follow-up imaging or biomarker measurements). Are the inferred trajectories biologically plausible? Do they correlate with actual clinical events? Beyond the lack of direct validation data, there is a structural reason why biological validity cannot be confirmed from the current experimental design: DyPro's residual evolution module is trained entirely through outcome supervision — its loss function optimizes for prediction accuracy on observed DFS and OS events. There is no biological supervision signal in the training objective. As a result, the inferred trajectories are optimized to minimize prediction error, not to faithfully represent underlying disease biology. A trajectory that improves

C-index may do so by exploiting statistical correlations in the MSKCC dataset that have no causal connection to actual tumor progression or liver regeneration dynamics. This means that even if serial follow-up imaging were available, a correlation between inferred trajectory shape and observed imaging progression would not confirm biological fidelity — it could simply reflect shared reliance on correlated prognostic features. True biological validation would require prospective studies with mechanistic endpoints, which are beyond the scope of the current evidence base. What experiment would reduce this uncertainty: retrospective analysis comparing inferred trajectory shapes with actual follow-up imaging progression patterns in patients with available serial CT scans, interpreted with appropriate caution about the outcome-supervision confound.

3. **Per-patient trajectory interpretability is absent.** DyPro provides no mechanism to explain why a particular patient receives a particular risk score. TrajSurv demonstrates that trajectory-level interpretation (vector fields, archetype clustering) is achievable. How would this be implemented in DyPro? Which residual evolution step (year 1 vs. year 5) most drives the risk prediction for a given patient?

4. **The optimal trajectory discretization is untested.** DyPro uses T=12 yearly steps. TrajSurv's NCDE approach (continuous-time) suggests finer discretization may capture more signal. Does DyPro's performance improve with 6-month or quarterly steps? Is there a diminishing returns threshold beyond annual discretization?

5. **The dependency on radiologist-annotated segmentations limits scalability.** DyPro requires precise multi-region CT segmentation. Automated segmentation accuracy varies across scanners and institutions. The Frontiers external validation finding — that scanner-dependent shape features drive the biggest performance drop — directly threatens DyPro's deployment if automated segmentation quality is inconsistent.

6. **Comparison with 4D-ACFNet is absent.** The same research group produced two architecturally distinct approaches to the same problem on the same dataset. No head-to-head comparison exists. Which approach is more robust to scanner variation? Which generalizes better? Which is more computationally efficient for clinical deployment?

7. **Decision-curve analysis has not been conducted.** DyPro's C-index advantage over CRS (+0.23) is substantial, but whether this translates into clinically net benefit (different treatment decisions that improve outcomes) is unknown. Decision-curve analysis would quantify the net benefit of DyPro-guided decisions vs. treat-all or treat-none strategies across different risk thresholds.

7. Recommended Next Steps

1. **Conduct external validation on an independent CRLM cohort.** This is the single most important next step. Obtain preoperative CT and clinical data from a CRLM cohort at a different institution. Apply DyPro's trained model (with frozen weights) to this external cohort and evaluate C-index, AUC, and IBS. This directly addresses the most critical unresolved question. The expected outcome based on prior CRLM radiomics literature is a performance drop of 5-15% in C-index, but if DyPro's peritumoral anatomical nodes provide robustness, the drop may be smaller than observed for simpler models.

2. **Implement and validate per-patient trajectory interpretability.** Using the trained DyPro model, extract the 12-step latent trajectory for each patient in the MSKCC validation set. Apply k-means clustering to the trajectory embeddings (following TrajSurv's approach) to identify 3-5 progression archetypes. Evaluate whether archetype membership correlates with actual survival outcomes (log-rank test). This addresses DyPro's interpretability gap and could reveal clinically meaningful patient subgroups.

3. **Run a head-to-head comparison with 4D-ACFNet on the MSKCC dataset.** Evaluate both methods using identical cross-validation splits and metrics. Assess whether the performance difference (if any) is consistent across patient subgroups stratified by tumor burden, surgical complexity, or clinical risk score. This comparison would determine whether the graph-based approach (DyPro) or the attention-based approach (4D-ACFNet) is superior for CRLM prognosis and guide future development direction.

4. **Replace manual segmentations with automated segmentation and assess performance impact.** Apply TotalSegmentator (nnU-Net backbone) or a comparable state-of-the-art abdominal CT automated segmentation tool to generate anatomical region segmentations for the MSKCC dataset. Target Dice coefficient ≥ 0.85 for major anatomical regions (liver, FLR, hepatic veins, portal veins) as the acceptance threshold for clinical-grade segmentation quality. Re-evaluate DyPro's C-index with automated vs. manual segmentations to quantify the segmentation quality dependency. This directly addresses the practical deployment scalability constraint.

5. **Conduct decision-curve analysis.** Using the MSKCC validation results, compute net benefit at clinically relevant risk thresholds (e.g., 20%, 30%, 50% 5-year recurrence probability). Compare DyPro-guided decisions against CRS-guided decisions and a treat-all strategy. This quantifies clinical utility beyond statistical discrimination metrics.

6. **Ablate peritumoral vs. intratumoral node contributions.** Systematically remove anatomical node types (keeping only metastatic tumor node vs. keeping all except tumor node) to quantify the contribution of peritumoral context to both discrimination and calibration. If peritumoral nodes are found to contribute disproportionately to calibration (vs. discrimination), this would

confirm them as the key generalization mechanism and prioritize their inclusion in any external validation.

7. **Investigate finer trajectory discretization.** Implement DyPro variants with T=24 (6-month steps) and T=48 (3-month steps) and evaluate whether additional temporal resolution improves C-index or IBS on MSKCC. This directly tests whether DyPro's performance is limited by its annual discretization.

8. Key Risks, Caveats, and Evidence Boundaries

1. **Single-cohort evidence is a fundamental limitation.** DyPro's performance (C-index 0.755 OS, 0.714 DFS) is demonstrated on one cohort (MSKCC, 197 patients). The literature consistently shows that CRLM prognostic models trained on MSKCC do not fully generalize to independent cohorts, consistent with findings from Simsek et al. (Frontiers in Digital Health 2025) on the same MSKCC/TCIA CRLM dataset. This evidence boundary should be explicitly stated in any report based on DyPro's current results.

2. **The inferred trajectories are computationally derived, not biologically observed.** DyPro's latent residual evolution generates postoperative risk trajectories that are never validated against actual postoperative follow-up data. The trajectories are computationally plausible and improve prediction performance, but their biological validity — whether they correspond to actual disease progression patterns — is unconfirmed.

3. **The DFS→OS cascade benefit may be dataset-specific.** DyPro's DFS→OS cascade improvement (+0.030 C-index) is observed on MSKCC. The clinical dependency between recurrence and survival may vary across populations. External validation is needed to confirm whether the cascade benefit generalizes.

4. **GraphSAGE's inductive bias advantage may not transfer.** DyPro adopts GraphSAGE for its neighborhood sampling-and-aggregation mechanism. While this improves over GCN and GAT on MSKCC, the advantage may not hold on different patient populations with different graph structures.

5. **The model's dependency on specific CT protocols is uncharacterized.** DyPro uses contrast-enhanced CT scans from MSKCC with specific acquisition parameters. The model's sensitivity to contrast timing, slice thickness, and reconstruction kernel is unknown. Given the Frontiers finding that scanner-dependent features drive generalization failures, this risk is non-trivial.

6. **External validation studies may have used simpler models that are more robust to distributional shift.** The consistent finding that radiomics models trained on MSKCC degrade in external validation comes from studies using simpler models (radiomics signatures, traditional ML). If DyPro's additional complexity overfits to MSKCC-specific patterns more severely, its external validation drop could be larger than observed for simpler models.

This risk is elevated by DyPro's reliance on pretrained 3D ResNet18 features, which encode scanner-specific image statistics.

7. **Clinical workflow integration requires more than predictive performance.** DyPro provides risk scores but no decision support framework. Translating C-index improvements into clinical decisions (adjuvant therapy selection, follow-up frequency) requires decision-curve analysis, cost-effectiveness evaluation, and clinician acceptance testing — none of which have been conducted.