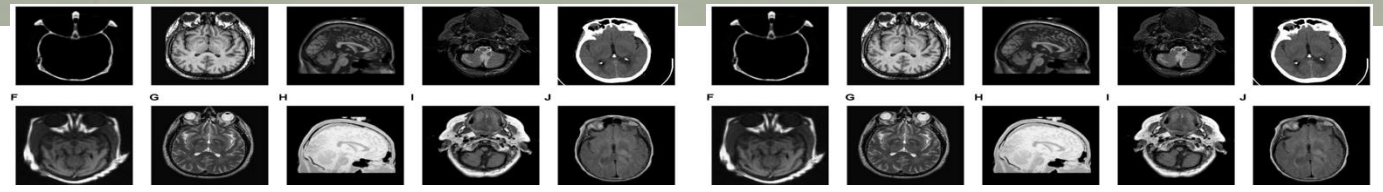
Multimodal Fusion for Disease Understanding

Tanveer Syeda-Mahmood

IBM Fellow, Adjunct Professor

Department of Biomedical Data Science, Stanford University



Stanford MEDICINE

Acknowledgements

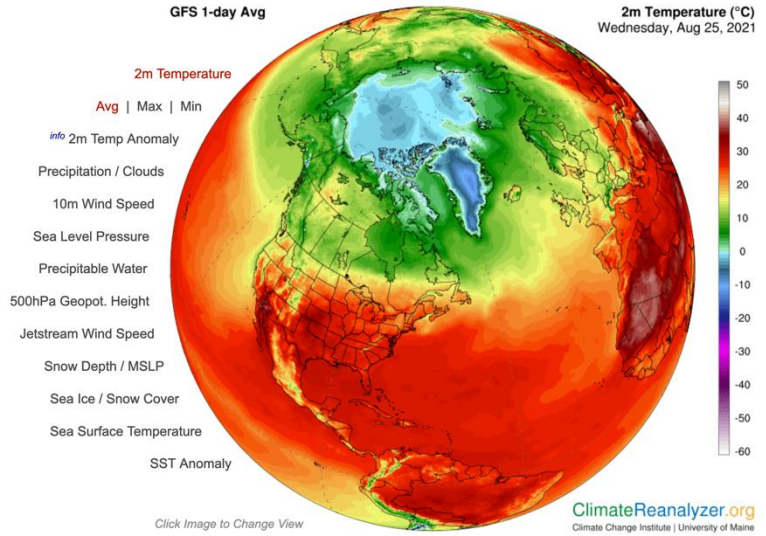
- Work represents research done in collaboration with
  - UIUC (Vaishnavi Subramanian, Minh Do)
  - University of Leeds (Rodrigo Benezola, Alex Frangi)
  - IBM Labs:
    - Niharika Dsouza (IBM Almaden)
    - Hongzhi Wang (IBM Almaden)
    - Ken Wong (IBM Almaden)
    - Antonio Foncubierta (IBM Zurich Labs)
    - Andrea Giovanni (IBM Zurich Labs)
    - Maria Gabrani (IBM Zurich Labs)



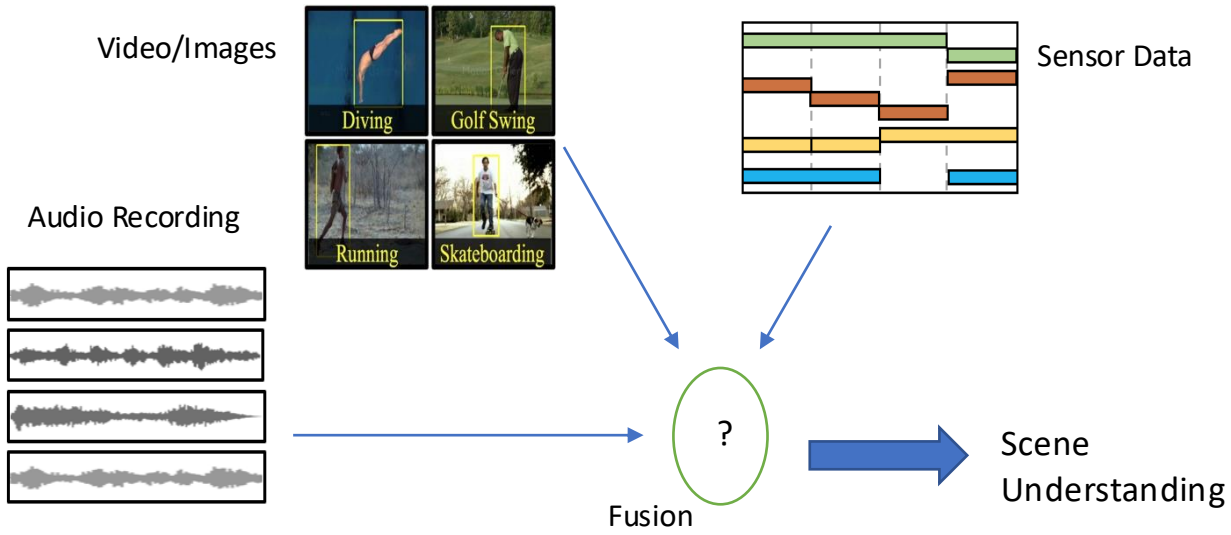



The Multimodal Fusion Problem

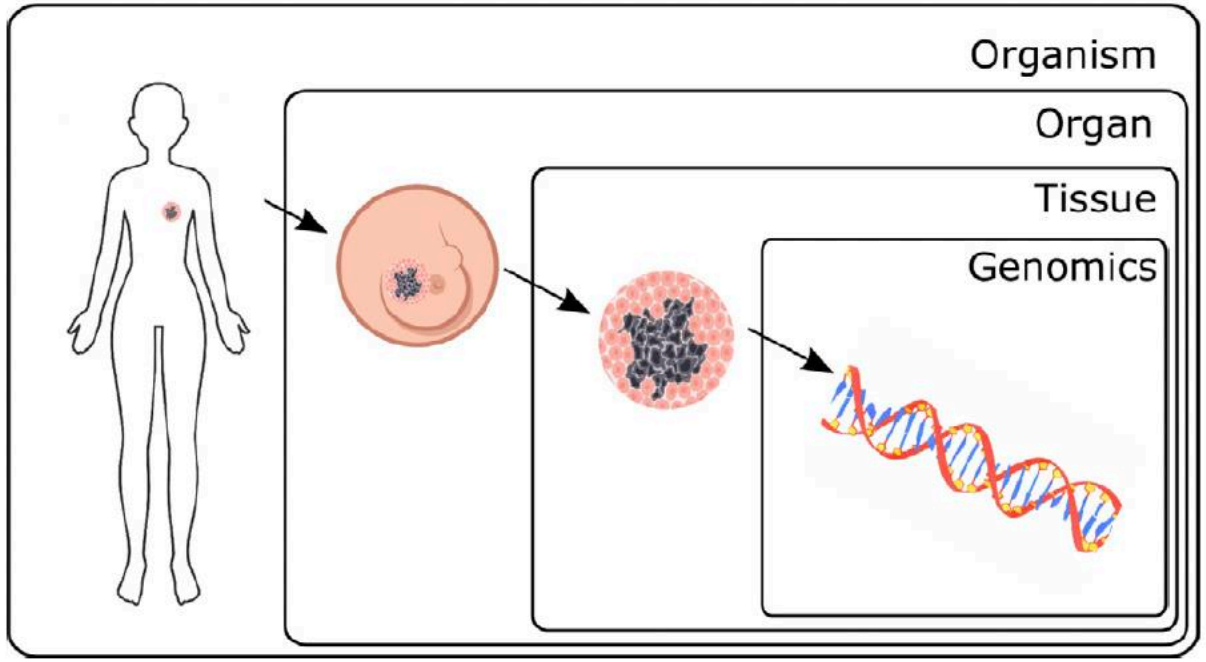
- Observations about a problem can be made through many modalities over different scales
  - **The number of observation modalities can be very large in some domains**
  - **The dimensions of the modalities can be large**

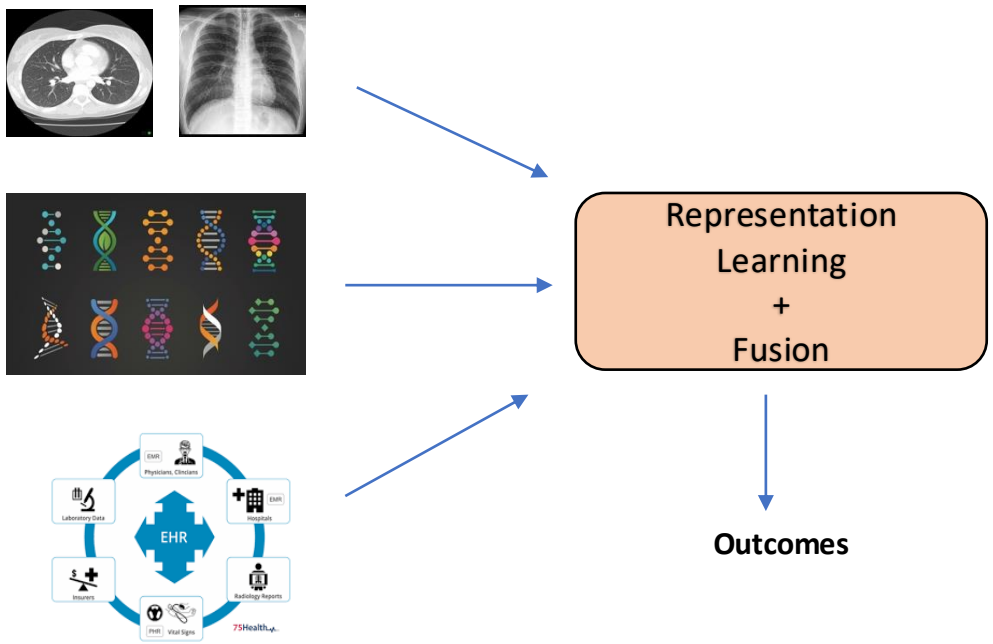


- Effectively leveraging all the information present in these modalities is important
  - Improved detection
  - Better prediction and recurrence
  - Understanding the mechanisms
  - **Problem characterization**
  - **Reconstruction/regression/mapping estimation**
  - **Digital twin modeling**



Multimodal Fusion For Healthcare





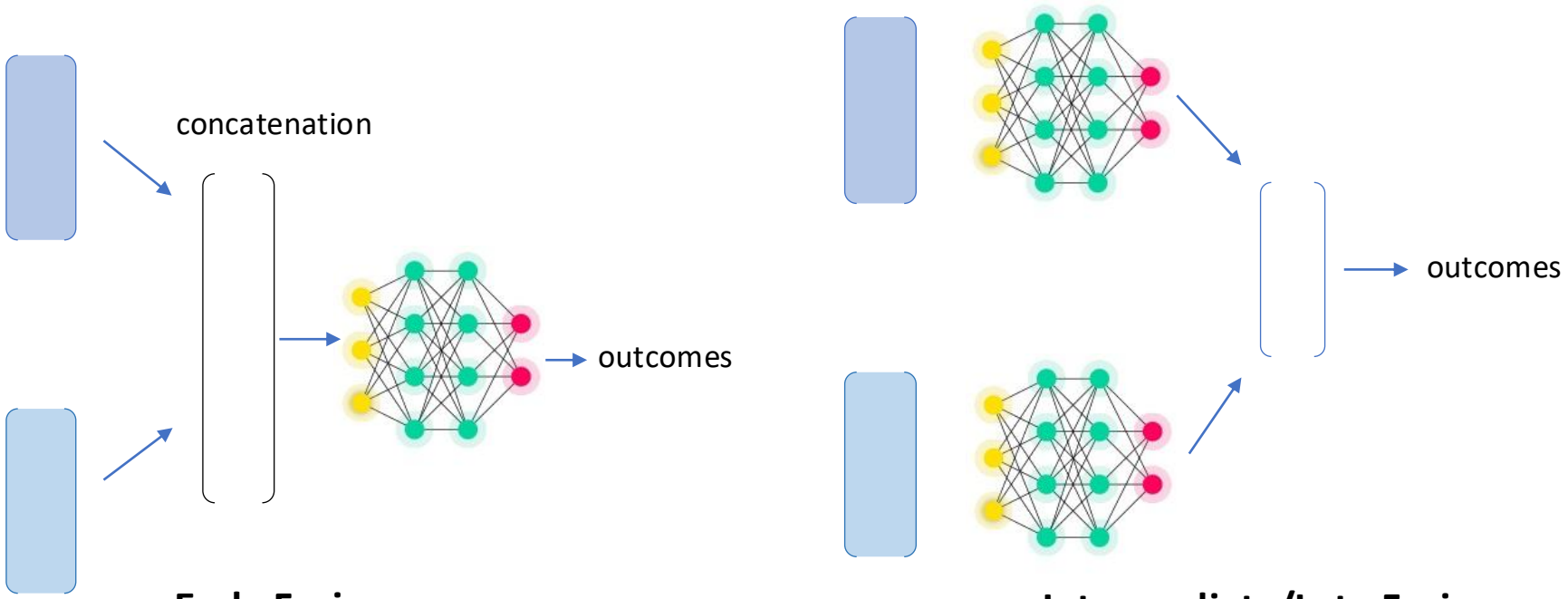
Challenges

- Inference from each modality is not definitive and may have uncertainty or errors
- Each modality may need a different representation to capture its information
- Not all modality evidence may be present or relevant for each instance of the problem
- Number of samples smaller than the dimensionality per modality

Multimodal fusion in healthcare - The challenges

- Representation
  - How to represent and summarize multimodal data across scales ?
  - Each modality may need a different representation to capture its information
- Alignment
  - How to capture the relationship between (sub)elements from two or more different modalities (spatial and temporal) across scales?
  - Capture modality relationships
- Fusion
  - How to combine information that is mutually reinforcing or complementary?
  - Fusion may need to be across both space and time
  - When does fusion help?
- Uncertainty handling
  - Inference from each modality is not definitive and may have uncertainty or errors
    - E.g. Evidence is pathology imaging assumes the tissue region was sampled correctly.
- Missing data
  - Not all modality evidence may be present for each instance of the problem
- Lack of coordinated datasets
  - No large-scale datasets that span all scales

Traditional Multimodal Fusion Strategies



Early Fusion **Intermediate/Late Fusion**

Limitations:

- Complexity grows with increasing number of concatenated features
- Learning and fusing representations at scales
- Exploiting correlations/inter-dependence within modality features

Can Foundational Models Help?

Transformers

- Your kitchen sink of good ideas all rolled into one architecture
  - Tokenizing the input text
  - Positional encoding of tokens to preserve order
  - Attention – the Q,K,V paradigm
  - Multi-head attention – for more combinations to explore
  - Multi-layer encoders for increased abstractions
  - Linear layers for aggregating it all together
  - Explores this in parallel for all tokens unlike RNN
  - Representations created more powerful than CNNs!

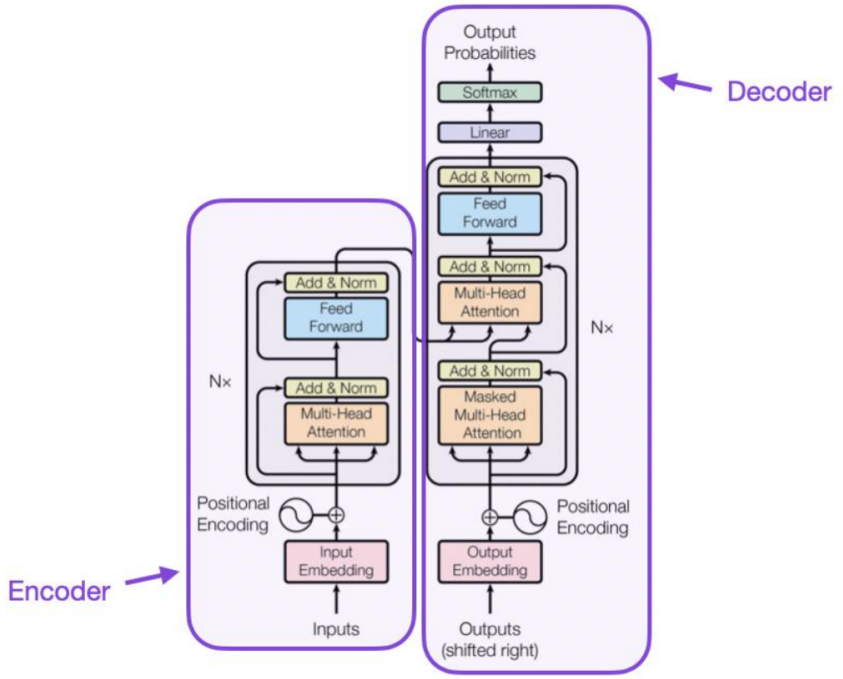


Figure 1: The Transformer - model architecture.

https://www.linkedin.com/pulse/llm-transformer-architecture-shivasish-mahapatra-kj9qf/

Fusion methods in AI models

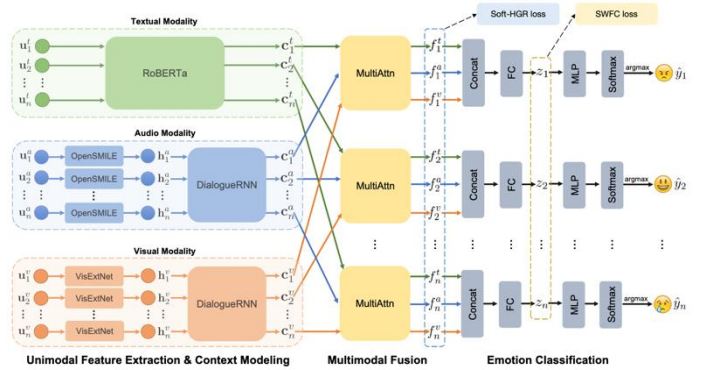
- Early fusion:
  - combines raw data or low-level features from different modalities at the input stage, before significant processing by the main model.
  - **Benefits:**
    - This approach can capture interactions between modalities early in the process, leading to a richer representation.
  - **Drawbacks:**
    - careful alignment and synchronization of data and can lead to high-dimensional feature spaces, increasing computational complexity.
  - **Best Use Cases:**
    - It is suited for tightly coupled modalities where the interactions are crucial from the start, such as multimodal sentiment analysis combining facial expressions and speech tone.

Fusion methods in AI models

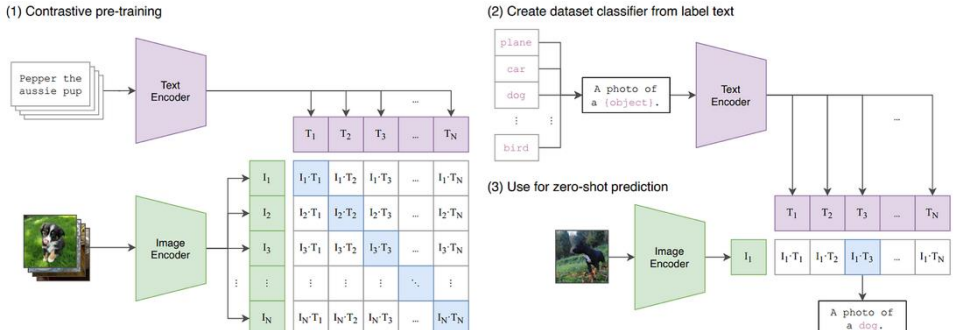
- Late fusion:
  - Process each modality separately
  - Fuse results by averaging, voting, or a meta-classifier to produce the final result.
  - MOE models
  - **Benefits:**
    - This method offers modularity and flexibility; individual models can be optimized independently and it is robust to missing modalities.
  - **Drawbacks:**
    - It may miss opportunities to model cross-modal interactions during the processing phases.
  - **Best Use Cases:**
    - It is suitable for scenarios where modalities are loosely related or asynchronous, such as combining patient medical images and textual records for a final diagnosis.

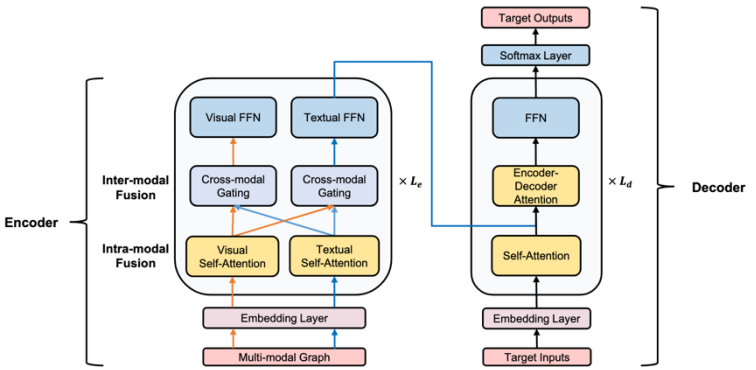
Fusion methods in AI models

- Intermediate or advanced fusion methods:
  - **Cross-Modal Attention:**
    - This is a key technique in modern models such as GPT-4V and BLIP-2. Attention mechanisms dynamically weigh the importance of features from one modality (e.g., a text prompt) while processing another (e.g., an image), allowing for nuanced and context-aware interactions.
  - **Joint Embedding Spaces:**
    - Many modern models, such as OpenAI's CLIP, project different modalities into a shared embedding space. In this space, the semantic similarity between different data types can be directly measured (e.g., an image of a cat and the word "cat" are close together), facilitating tasks like image retrieval or zero-shot classification.
  - **Encoder-Decoder Architectures:**
    - These models use separate encoders for each modality and a shared decoder (often a large language model) to generate the final output. The fusion typically happens within the connections between the encoders and the decoder



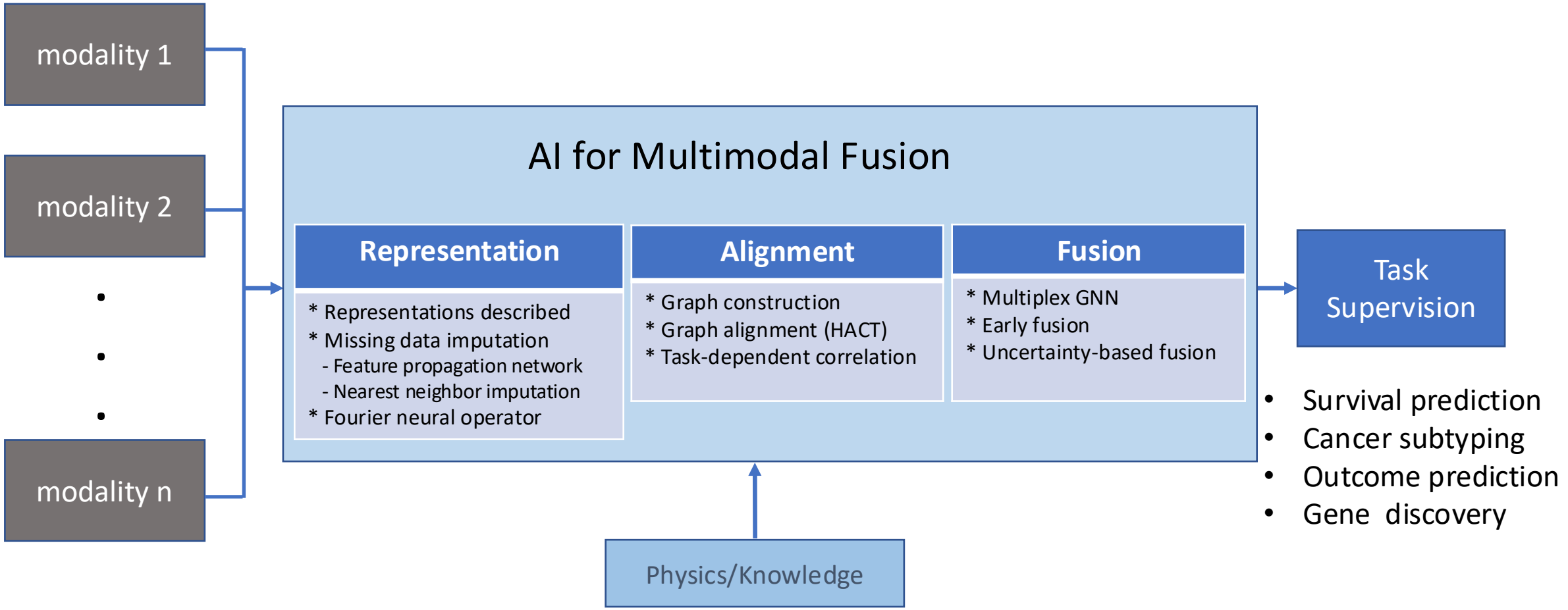
MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations, ACL 2023





A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation, ACL 2020

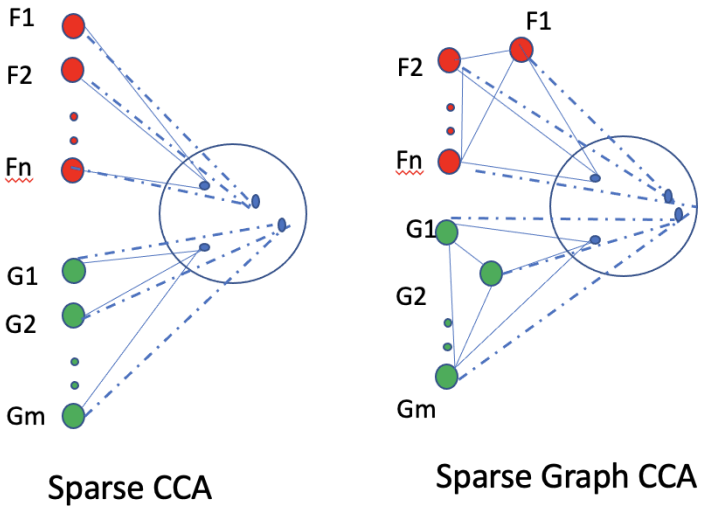
The Multimodal Fusion Problem – Our approach



Representative problems in multimodal fusion

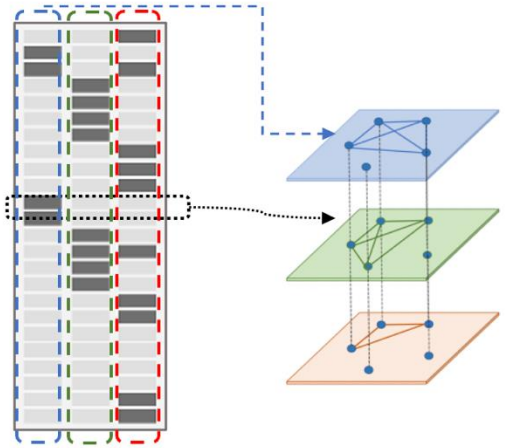
Feature Representations

Sparse graph CCA encoding (Vaishnavi et al ISBI 2021)



Multimodal Fusion

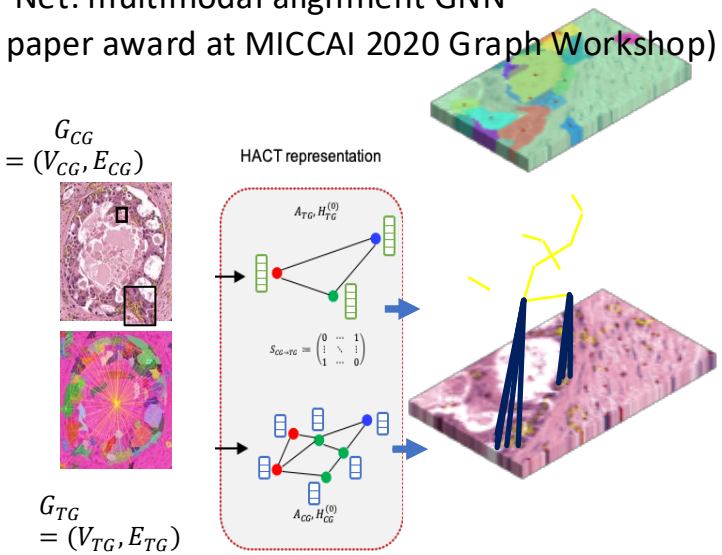
MICCAI 2022, MedIA'24, Young Scientist Finalist



Multimodal Alignment

HACT-Net: multimodal alignment GNN

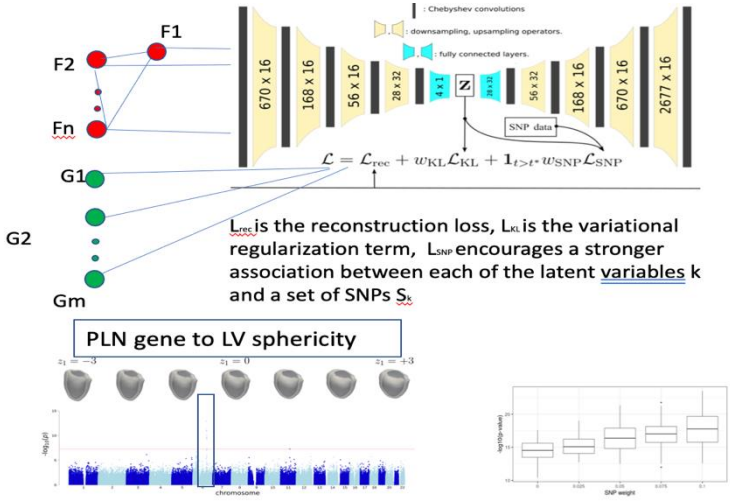
(Best paper award at MICCAI 2020 Graph Workshop)

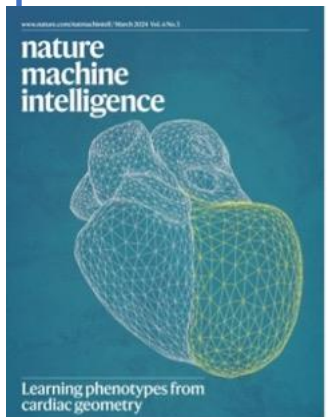


Multimodal Fusion for Discovery

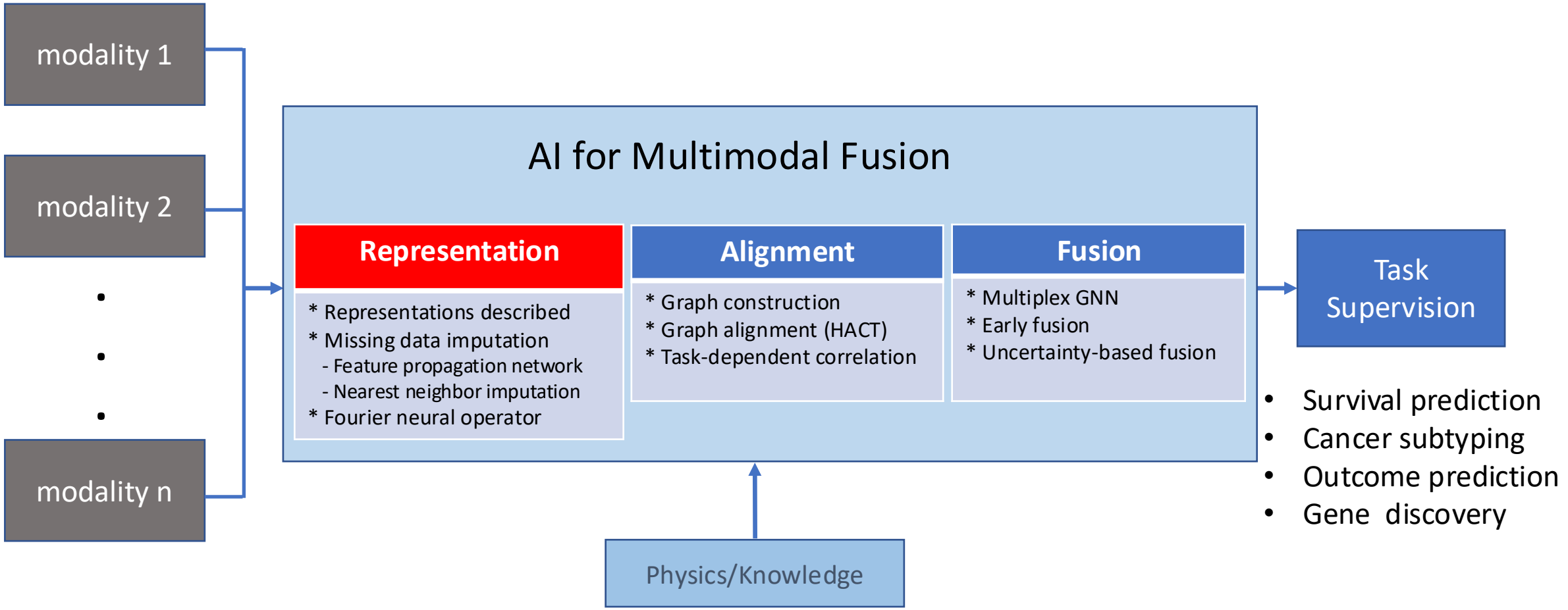
Cross-modal association discovery

(Rodrigues et al. MICCAI 2021)

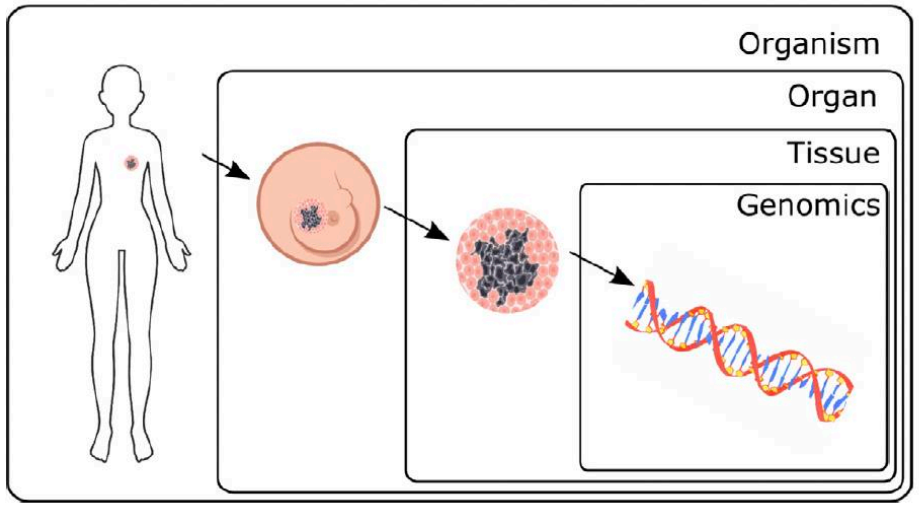




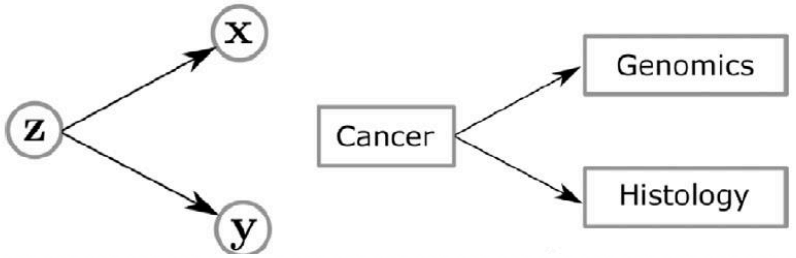
The Multimodal Fusion Problem – Our approach



Multimodal Fusion – Taking a modeling viewpoint



Probabilistic graphical model



$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \ \min\{p, q\} > d > 1,$$

$$\mathbf{x} \sim \mathcal{N}(\mathbf{W}_x \mathbf{z}, \boldsymbol{\Psi}_x), \ \mathbf{W}_x \in \mathbb{R}^{p \times d}, \boldsymbol{\Psi}_x \gneqq 0, \boldsymbol{\Psi}_x \in \mathbb{R}^{p \times p},$$

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}_y \mathbf{z}, \boldsymbol{\Psi}_y), \ \mathbf{W}_y \in \mathbb{R}^{q \times d}, \boldsymbol{\Psi}_y \gneqq 0, \boldsymbol{\Psi}_y \in \mathbb{R}^{q \times q},$$

Modalities usually originate from the same source (e.g. cancer site):

histology tissue imaging

radiology imaging

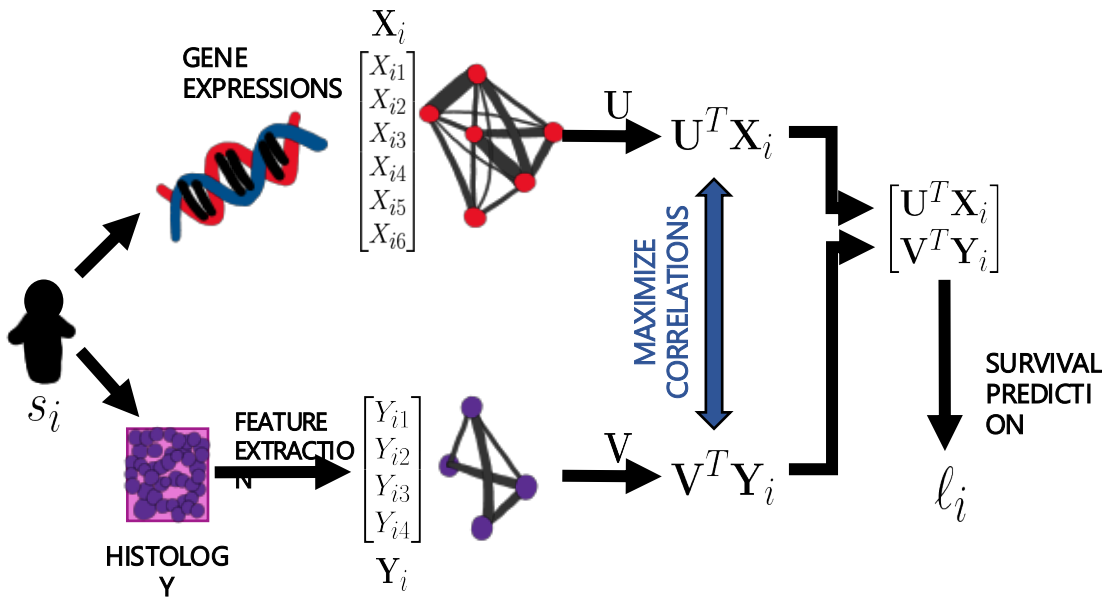
different levels of genomics (gene expression and transcriptomics)

Modalities account for the disease at different scales of observation

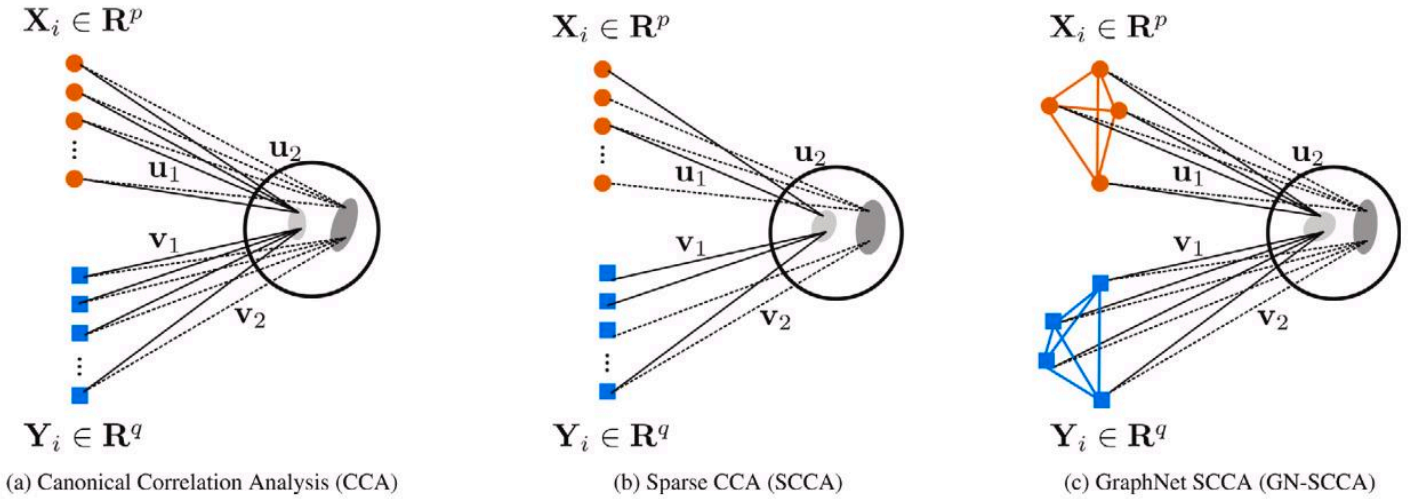
A probabilistic graphical model for fusion:

Estimate an underlying property of the disease, e.g. survival potential as a latent variable z, from observable modalities x,y,…

Representations learning



- Explore modality correlations:
  - Canonical correlation analysis (CCA)
  - HGR correlation
- Project using these embeddings
- Predict the task:
  - Survival prediction

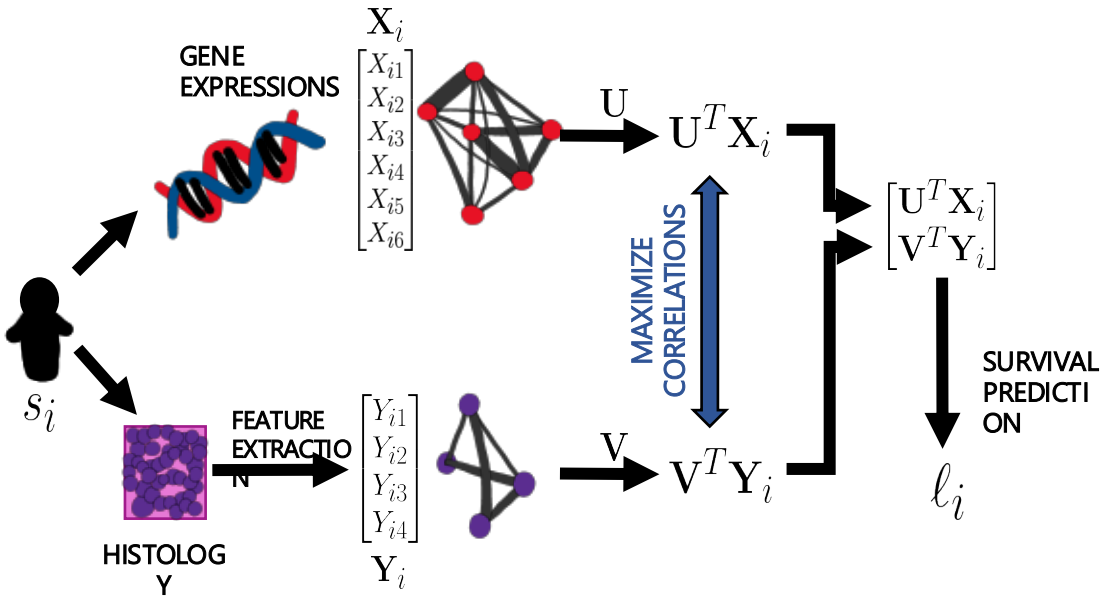


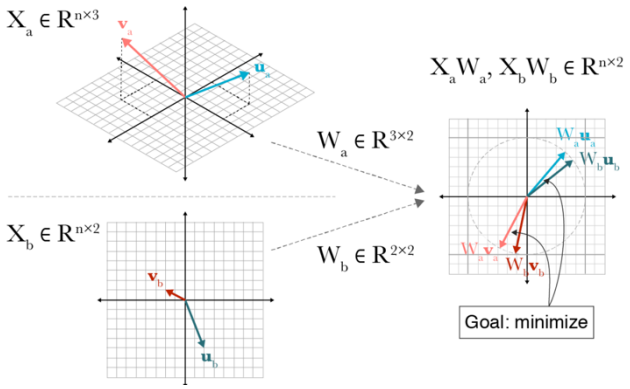
(a) Canonical Correlation Analysis (CCA)

(b) Sparse CCA (SCCA)

(c) GraphNet SCCA (GN-SCCA)

Representations learning





Canonical correlation analysis

$$\rho^* = \max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{C}_{xy} \mathbf{v} \quad \text{s.t.} \quad \mathbf{u}^T \mathbf{C}_{xx} \mathbf{u} = 1, \mathbf{v}^T \mathbf{C}_{yy} \mathbf{v} = 1,$$

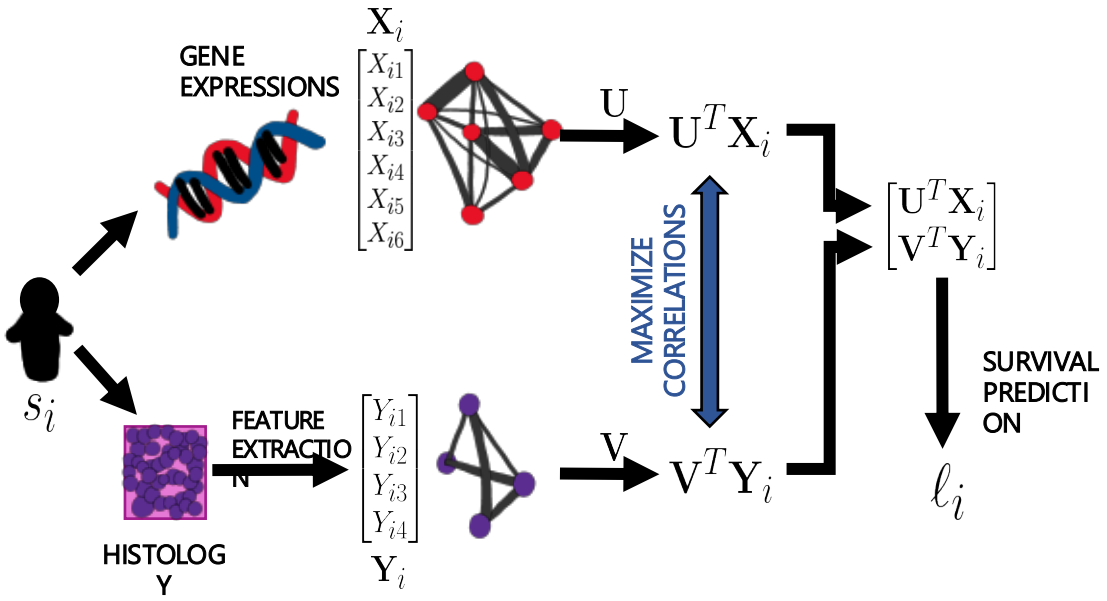
with optimal canonical weights $(\mathbf{u}^*, \mathbf{v}^*)$. The corresponding 1- dimensional CCA embeddings are $(\mathbf{u}^*)^T\mathbf{X}$ and $(\mathbf{v}^*)^T\mathbf{Y}$. This problem can be equivalently posed based on the singular value decomposition (SVD) as

$$\rho^* = \max_{\tilde{\mathbf{u}},\tilde{\mathbf{v}}} \tilde{\mathbf{u}}^T \mathbf{A} \tilde{\mathbf{v}} \quad \text{s.t.} \quad \tilde{\mathbf{u}}^T \tilde{\mathbf{u}} = 1, \ \tilde{\mathbf{v}}^T \tilde{\mathbf{v}} = 1,$$

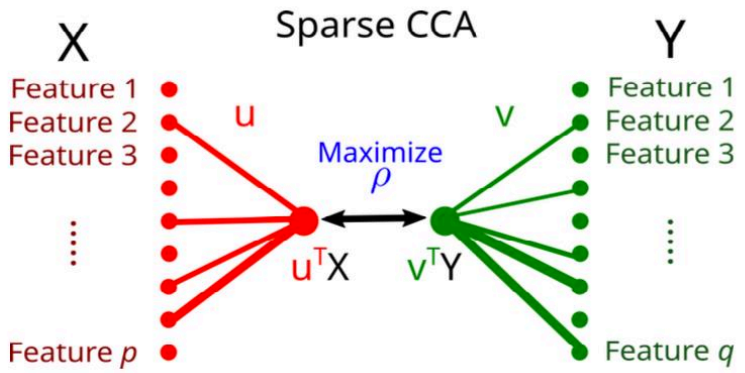
with $\mathbf{A} = \mathbf{C}_{xx}^{-1/2} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1/2}$, assuming invertibility of the auto-correlation matrices $\mathbf{C}_{xx}, \mathbf{C}_{yy}$. The maximizing arguments are the first singular vectors $(\tilde{\mathbf{u}}^*, \tilde{\mathbf{v}}^*)$ using the singular value decomposition (SVD) of $\mathbf{A}$ [34]. The optimal CCA canonical weights $\mathbf{u}^*$ and $\mathbf{v}^*$ can then be obtained as $\mathbf{u}^* = \mathbf{C}_{xx}^{-1/2} \tilde{\mathbf{u}}^*$, $\mathbf{v}^* = \mathbf{C}_{yy}^{-1/2} \tilde{\mathbf{v}}^*$.

Needs large sample size to get a solution via SVD

Representations learning



Sparse CCA



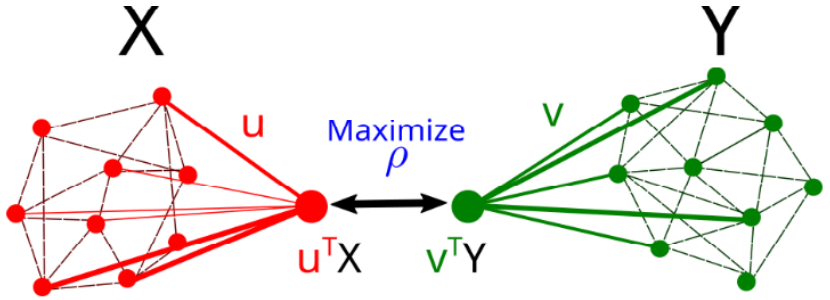
$$\rho^* = \max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T \mathbf{C}_{xy} \mathbf{v}$$

$$\text{s.t. } \|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1, \|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2,$$

Adding convex penalty constraints on the canonical weights

Graph-based Sparse CCA

- How to account for more general structural dependencies in features within a modality?
- What if structure is not known before-hand?
- Model the dependencies as a graph
- Use the Laplacian matrix to capture the correlation within modality
- Adds structure to complex bi-convex problem
- Enforce spectral frequency constraints



$$\hat{\mathbf{u}}, \hat{\mathbf{v}} = \arg\max_{\mathbf{u},\mathbf{v}} \ \mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}$$

$$\text{s.t. } \mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} \leq 1, \mathbf{v}^T\mathbf{Y}\mathbf{Y}^T\mathbf{v} \leq 1,$$

$$\|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2,$$

$$\mathbf{u}^T\mathbf{L}_1\mathbf{u} \leq d_1, \mathbf{v}^T\mathbf{L}_2\mathbf{v} \leq d_2.$$

$$\mathbf{u}^T\mathbf{L}\mathbf{u} = \sum_{(i,j)\in\mathcal{E}} w_{ij}(u_i - u_j)^2$$

Solving the biconvex optimization problem

$$\hat{\mathbf{u}}, \hat{\mathbf{v}} = \arg\max_{\mathbf{u},\mathbf{v}} \ \mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}$$

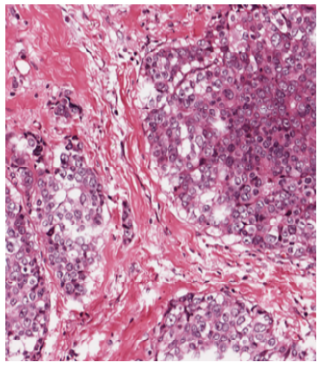
$$\text{s.t. } \mathbf{u}^T\mathbf{X}\mathbf{X}^T\mathbf{u} \leq 1, \mathbf{v}^T\mathbf{Y}\mathbf{Y}^T\mathbf{v} \leq 1, \ \|\mathbf{u}\|_1 \leq c_1, \|\mathbf{v}\|_1 \leq c_2, \ \mathbf{u}^T\mathbf{L}_1\mathbf{u} \leq d_1, \mathbf{v}^T\mathbf{L}_2\mathbf{v} \leq d_2.$$

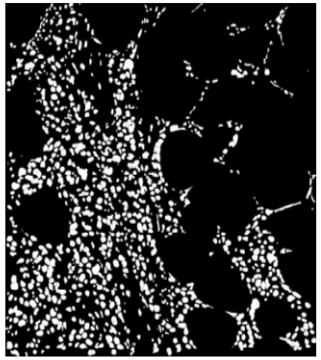
Iteratively solved using proximal gradient descent

Results

In ISBI 2021

Cancer Data - Feature Extraction | TCGA-BRCA: 973 patients





Genomics feature extraction ($\mathbf{X}$, $p = 500, 1000$, etc.)

- Most variant genes based on coefficient of variation ($\frac{\sigma}{\mu}$)
- Top $500, 1000$ genes' expression vales: genomic feature vector

Imaging feature extraction ($\mathbf{Y}$, $q = 213$)

- Randomly select 25 patches of size 4000 x 4000 pixels
- Nuclei segmentations from recent adversarial framework[14]
- Area and shape properties of nuclei extracted using CellProfiler[15]

Extraction of graphs

- Prior knowledge: Graph-Laplacians constructed from STRING-db
- Inferred from sample covariance matrices

[14] L. Hou et al. "Robust histopathology image analysis". In: *CVPR 2019*

[15] A. E. Carpenter et al. "CellProfiler: image analysis software". In: *Genome biology* (2006)

Results

- Graph-based Sparse CCA
  - common information between two modalities using linear projections
  - Handles correlations across features of the same-modality
  - A novel proximal-gradient optimization method for GCCA
  - Low-dimensional embeddings extracted meaningful correlations in genomic-pathology imaging for breast cancer
  - Used for 1 year survival prediction by using RF to classify the fused features.

Table 2: TCGA-BRCA: Estimated correlations in percentages (mean $\pm$ standard deviations) across 5 folds. Higher values desired.

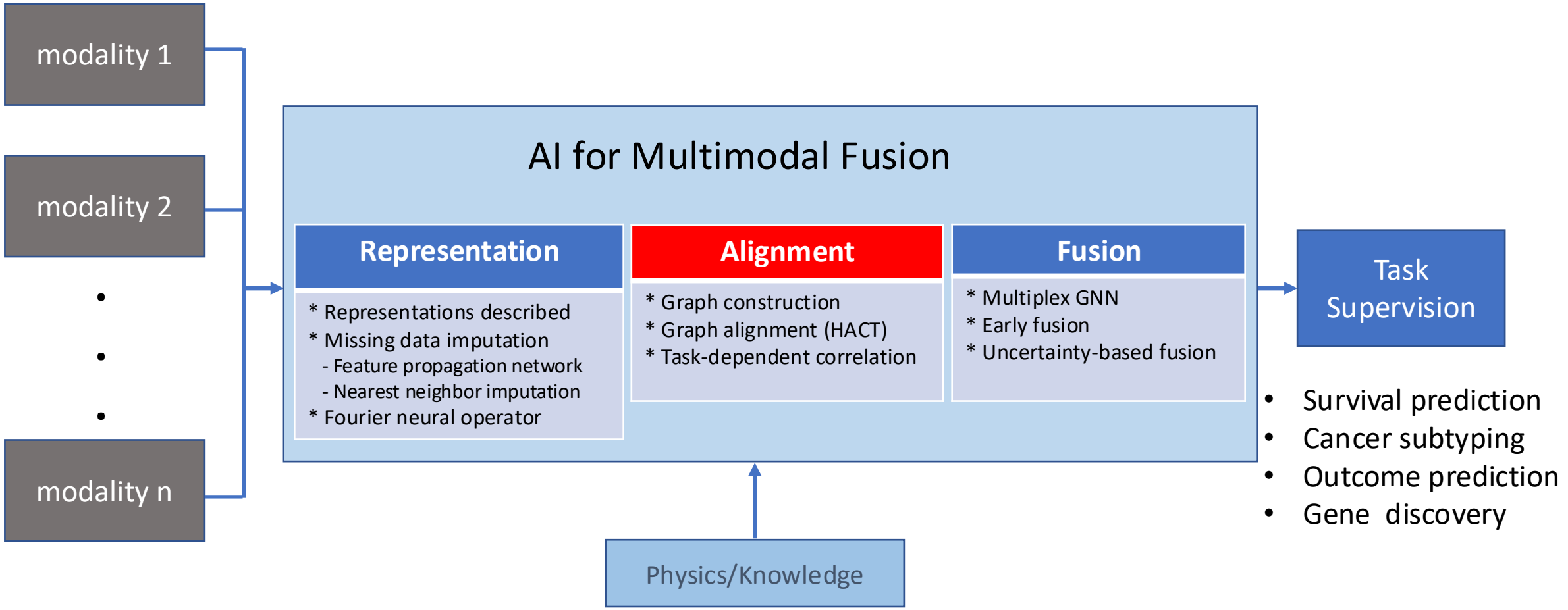
Method	$p = 100$	$p = 500$	$p = 800$	$p = 1000$	$p = 3000$
SCCA	19.33 ± 8.72	39.64 ± 3.91	42.77 ± 5.72	42.41 ± 5.35	45.60 ± 5.51
GN-SCCA	39.31 ± 11.63	51.44 ± 8.45	$\mathbf{58.29 \pm 6.80}$	58.47 ± 7.36	$\mathbf{63.24 \pm 6.55}$
GN-Prior	40.18 ± 6.91	52.25 ± 7.35	52.16 ± 5.93	53.02 ± 4.47	49.77 ± 4.81
GCCA (Ours)	44.36 ± 9.14	51.61 ± 9.33	54.97 ± 7.03	52.21 ± 8.72	48.46 ± 3.93
GCCA-Prior (Ours)	$\mathbf{46.94 \pm 3.92}$	$\mathbf{55.22 \pm 7.67}$	58.17 ± 5.10	$\mathbf{58.70 \pm 4.95}$	60.56 ± 5.31

A look at the top-weighted features across 5 folds for $p = 1000$ with GCCA:

Top genes	Function
FAM64A	PICALM interacting mitotic regulator
HJURP	Chromosome maintenance and chromatin regulation
RAD54L	Involved in DNA repair and mitotic recombination
GSG2	Serine/threonine-protein kinase that phosphorylates histone H3 during mitosis

Top-weighted Imaging Features
Nuclei Perimeter
Nuclei's Zernike Moment of Order (2,0)
Nuclei's Zernike Moment of Order (3,3)
Nuclei's Zernike Moment of Order (5,1)

The Multimodal Fusion Problem – Our approach



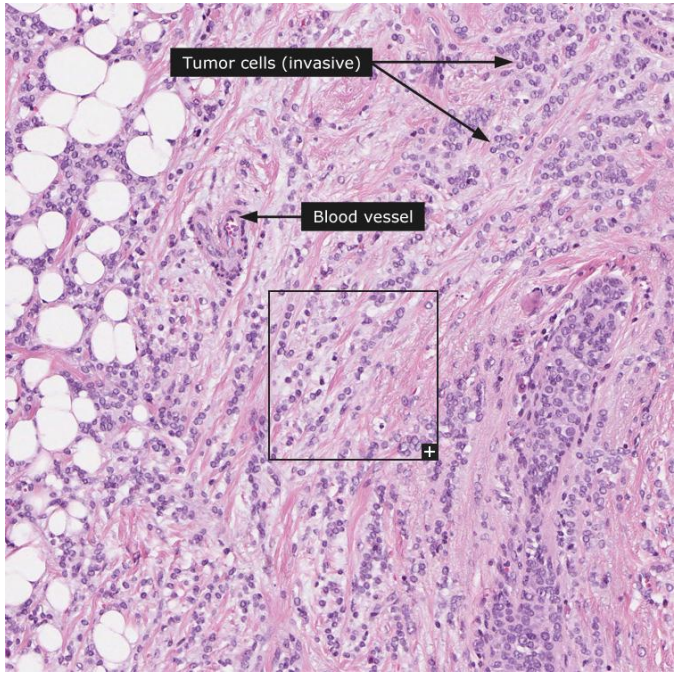
Multimodal alignment

- Alignment can be spatial or temporal or both
  - E.g. gene expression data and imaging data from the same tissue sample
- Alignment can be within a modalities, across modalities
  - Cell features and relations as graphs
  - Tissue features and relations as graphs
  - Cells to Tissue relationships
- Need deep learning architectures that can preserve spatial or temporal information.
  - Spatially-preserving flattening

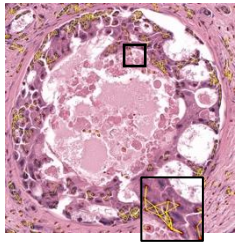
HACT defined as $G_{\text{HACT}} := \{G_{\text{CG}}, G_{\text{TG}}, S_{\text{CG}\to\text{TG}}\}$. G_{CG}

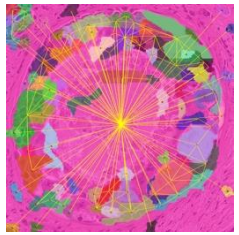
Breast cancer subtyping

- Normal, Benign, Atypical, DCIS, Invasive
- Modalities within digital pathology imaging
  - Cellular level
  - Tissue level

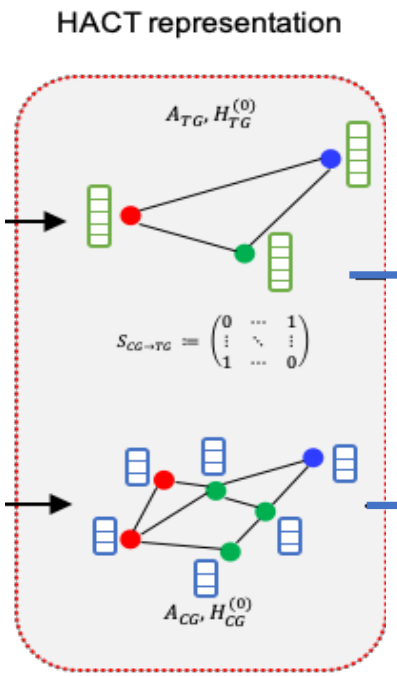


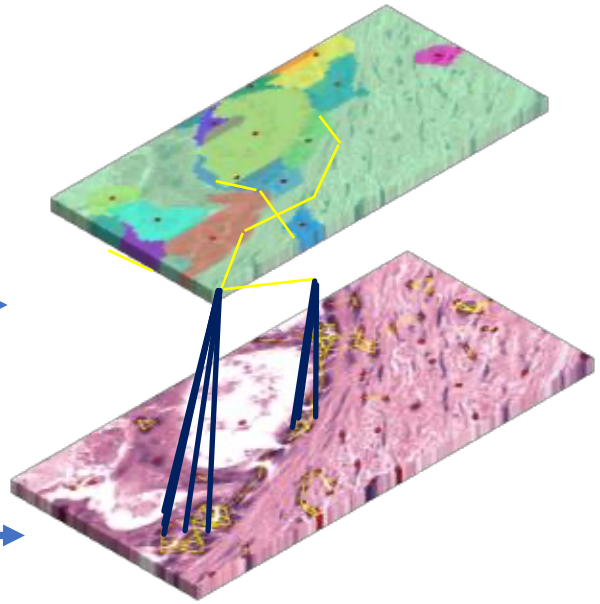
$G_{CG} = (V_{CG}, E_{CG})$





$G_{TG} = (V_{TG}, E_{TG})$





HACT-Net – A graph neural network-based fusion

In MICCAI 2020 Workshop

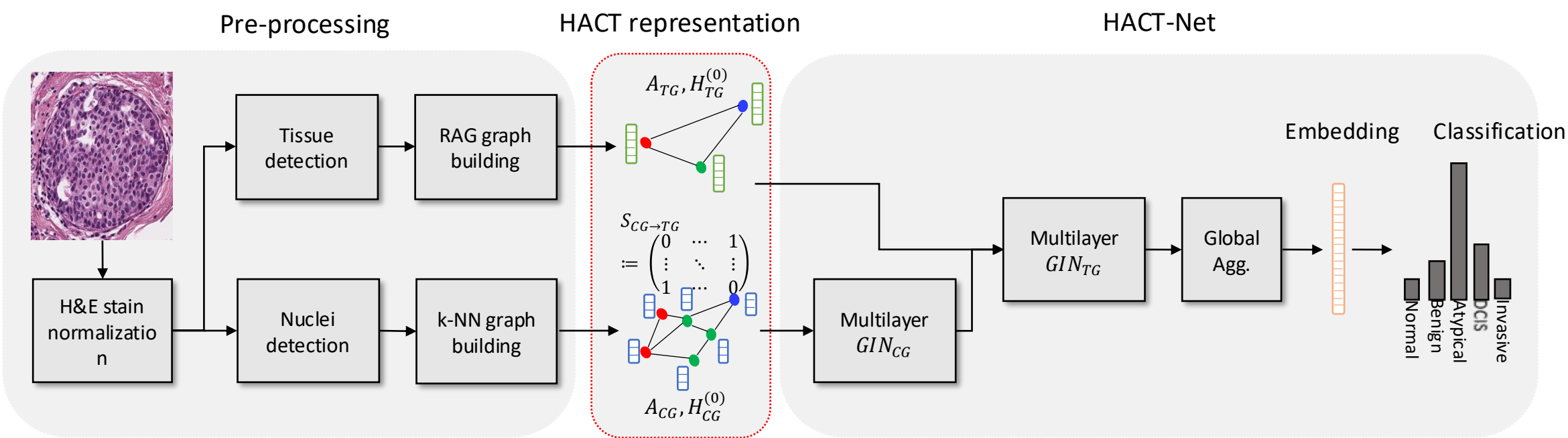
- Graph neural network
  - Each node *v* is characterized by its feature vector
  - Each edge represents a semantics and a weight
  - The state at each node is represented by an embedding

$$\mathbf{h}_v = f(\mathbf{x}_v, \mathbf{x}_{co[v]}, \mathbf{h}_{ne[v]}, \mathbf{x}_{ne[v]})$$

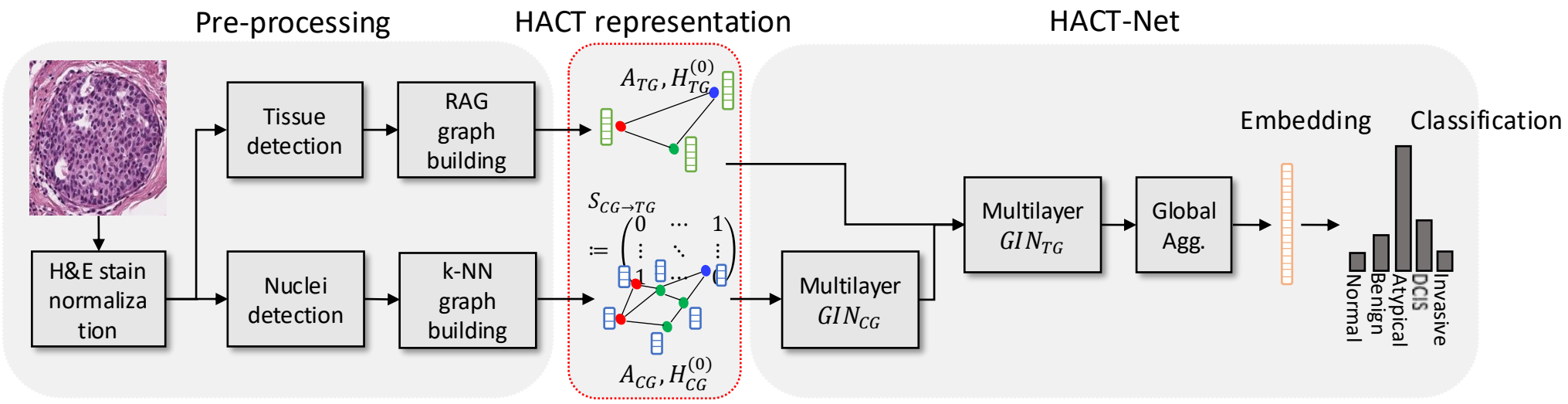
  - incorporates information about the node features, its edges, its neighbors, and their features
  - Embedding updated through a message passing iteration

$$\mathbf{H}^{t+1} = F(\mathbf{H}^t, \mathbf{X})$$

- Fusion method – HACT-Net
  - A hybrid fusion method
  - Graph neural network for each modality
  - Features per node derived separately as appropriate per modality
  - Elevate cell-level features through graph embedding
  - CG features selected for concatenation at each TG node come from the features of the cells that map to this tissue region using HACT



HACT-Net



$$h_{\mathrm{CG}}^{(t+1)}(u) = \mathrm{MLP}\left(h_{\mathrm{CG}}^{(t)}(u) + \sum_{w \in \mathcal{N}_{\mathrm{CG}}(u))} h_{\mathrm{CG}}^{(t)}(w)\right)$$

$$h_{\mathrm{TG}}^{(0)}(v) = \mathrm{Concat}\left(f_{\mathrm{TG}}(v), \sum_{u \in \mathcal{S}(v)} h_{\mathrm{CG}}^{(T_{\mathrm{CG}})}(u)\right) \quad \mathcal{S}(v) := \{u \in V_{\mathrm{CG}} \mid S_{\mathrm{CG}\to\mathrm{TG}}(u, v) = 1\}$$

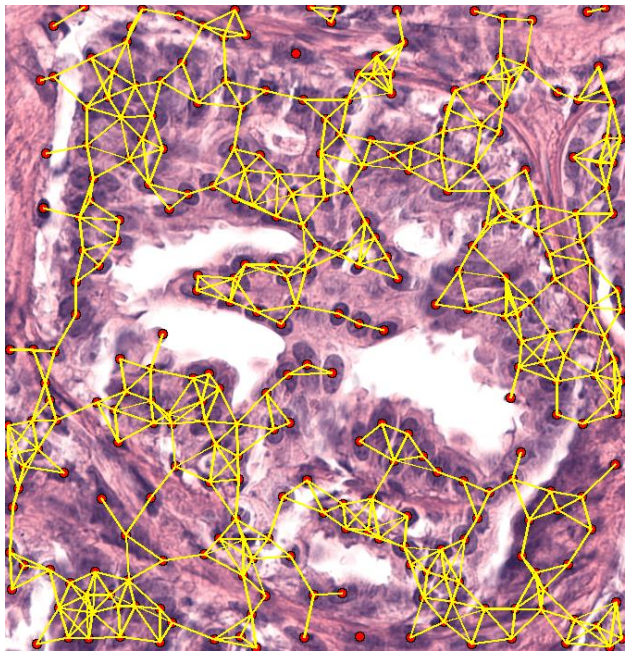
$$h_{\mathrm{HACT}} = \mathrm{Concat}\left(\left\{ \sum_{v \in G_{\mathrm{TG}}} h_{\mathrm{TG}}^{(t)}(v) \;\middle|\; t = 0, \ldots, T_{\mathrm{TG}} \right\}\right)$$

Results

Dataset: BRACS dataset for BReAst Carcinoma Subtyping (BRACS) . 2080 RoIs acquired from 106 H&E stained breast carcinoma whole-slide-images (WSI)
Data acquired through a JSA with PASCALE (Italy)

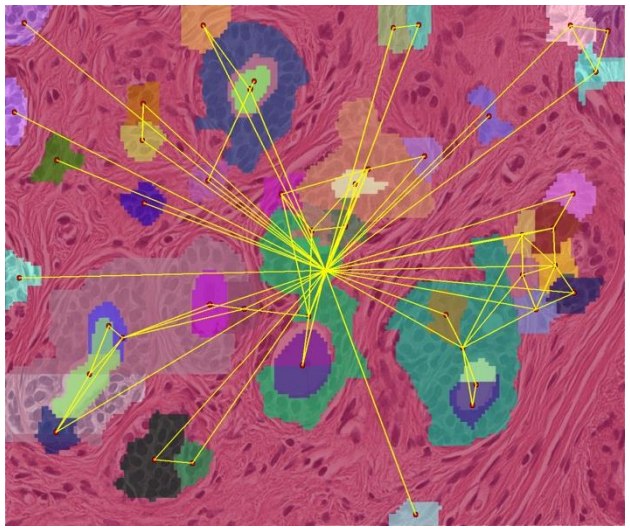
Cell level

Model/Fold#	1	2	3	4	$\mu \pm \sigma$	Normal	Benign	Atypical	DCIS	Invasive
CNN (20×)	59.12	58.38	55.33	58.26	57.77 ±1.45	62.00 ±3.67	52.00 ±4.30	46.50 ±5.22	59.75 ±4.92	70.00 ±3.94
CNN (40×)	55.21	53.38	47.97	57.20	53.44 ±3.44	64.75 ±8.17	50.75 ±4.71	39.00 ±4.30	50.25 ±3.90	67.25 ±4.49
TG-GNN	54.47	55.13	67.84	49.85	56.82 ±6.67	56.78 ±1.89	54.76 ±6.62	48.52 ±8.76	56.53 ±12.78	69.52 ±11.00
CG-GNN	61.35	53.81	62.00	55.38	58.13 ±3.59	62.66 ±5.32	**64.57** ±9.05	36.18 ±6.85	59.98 ±1.43	68.12 ±2.52
Concat-GNN	54.66	54.49	64.59	**63.95**	59.42 ±4.85	57.00 ±4.06	60.31 ±8.36	49.62 ±4.71	60.65 ±4.94	68.94 ±12.47
HACT-Net	**62.17**	**59.06**	**69.41**	60.92	**62.89** ±3.92	**65.15** ±3.64	58.40 ±10.59	**55.45** ±5.19	**63.15** ±4.08	**73.78** ±7.35

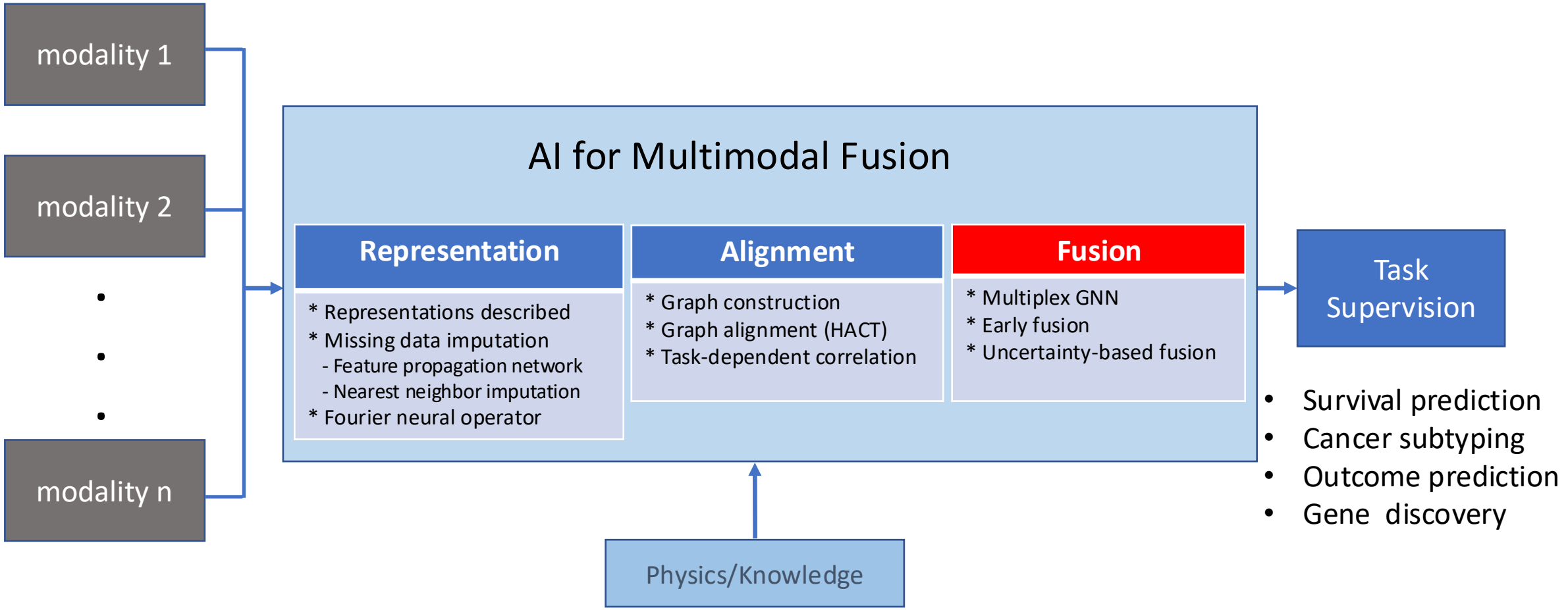


Best Paper Award at the MICCAI Workshop on Graphs

Tissue level



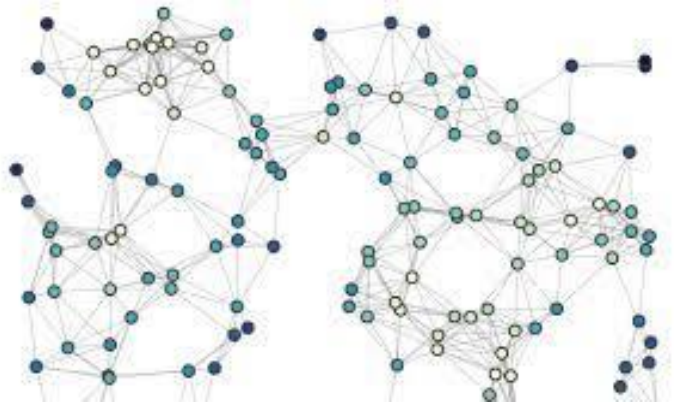
The Multimodal Fusion Problem – Our approach



Modeling Multimodal Fusion through Graphs

- Multimodal data can be modeled as a graph, where each data object is regarded as a node, and both intra- and inter-modal dependencies existing between data objects can be regarded as edges.
- Aggregation of information through message passing for downstream task such as disease or outcome prediction
- Key problems:
  - Constructing the graph
    - As the graph structure is not part of a dataset of patient information, the crucial point of GNN-assisted disease prediction is the graph creation.
      - Graphs constructed through non-imaging features
      - Based on similarities in latent representations
      - Adaptive graph construction based on fused multimodal features, dynamically adjusted during model training.
      - Multimodal embeddings and the original feature values of patients through a weighted summation
      - Sparsity filtering
        - degree-sensitive edge pruning and kNN sparsification strategies to eliminate redundant and noisy edges during graph creation and merging.
  - Propagation through message passing
    - A form of representation learning
      - Allow granular interactions within and across different modalities to learn more robust data representations for multimodal
    - Graph interaction networks – model the relationships
    - Graph attention networks to capture intra-modal contextual information and inter-modal complementary information.
    - Graph isomorphism networks (GIN) – captures distinct graph structures

Graph Based Representation Learning



$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),\ |\mathcal{V}| = P$$

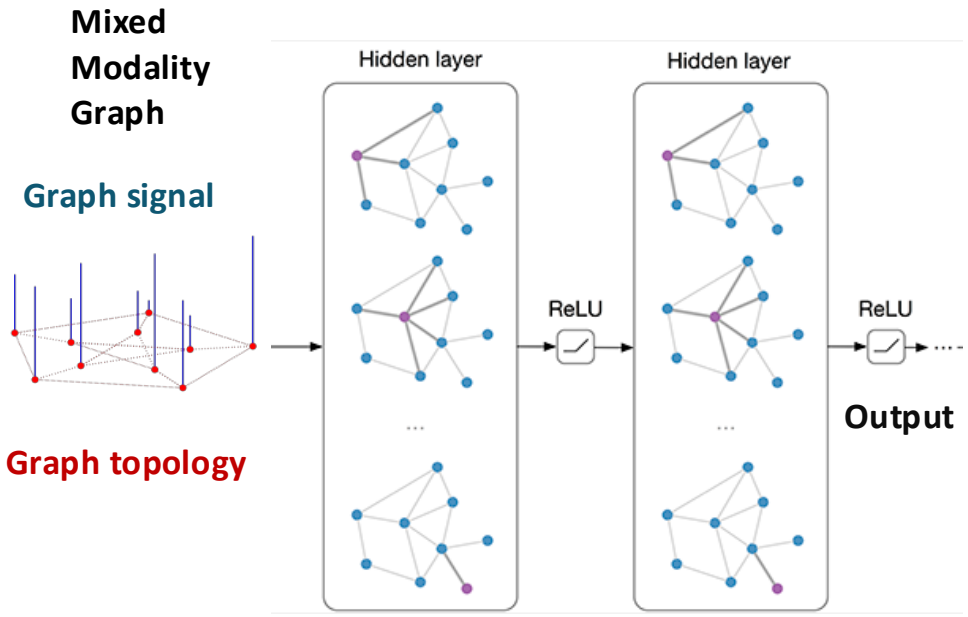
Nodes: Features/ Entities
Edges: Capture dependence within features

Adjacency Matrix: $\mathbf{A} \in \mathcal{R}^{P \times P}$

$$\text{If } (i,j) \in \mathcal{E}, \text{ then } \mathbf{A}(i,j) = 1$$
$$\text{else } \mathbf{A}(i,j) = 0$$

- Do not require spatial contiguity/organization of nodes
- Can represent data at arbitrary scales

Graph Neural Networks



Basic Message Passing

$$\mathbf{H}_i^{(l+1)} = \sigma\Big[\sum_{j \in N(i)} \mathbf{H}_j^{(l)} \mathbf{W}^{(l)}\Big]$$

non. linearity

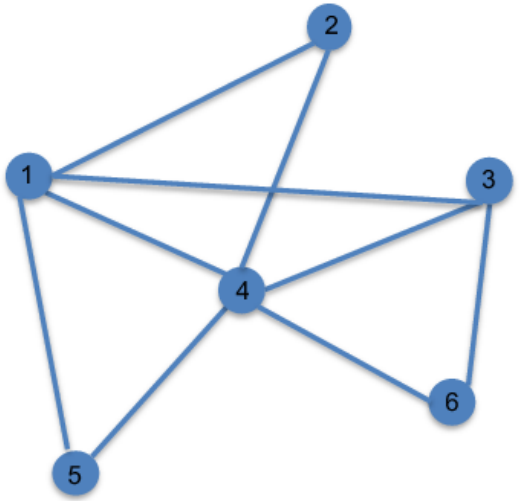
Messages at layer l

Filter at layer l

Graph topology: Intrinsic connectivity in the graph
Graph signal: Node/edge attributes for propagation

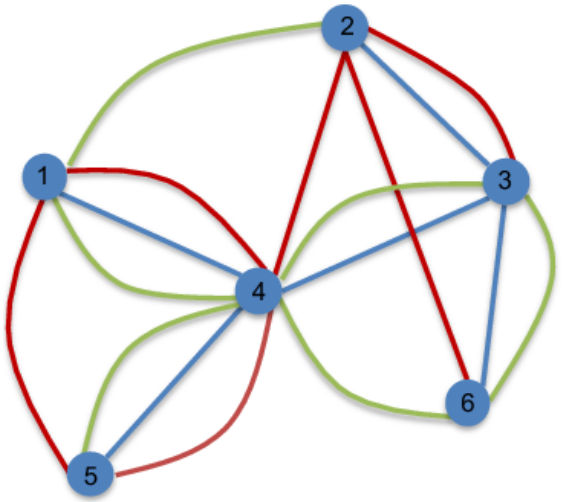
Message Passing: Tracking the information flow within the graph based on local neighbourhoods

Multi-Graphs for Multimodal Learning



Single view graph:
Graph with single edge types

$$\mathbf{A} \in \mathcal{R}^{P \times P}$$



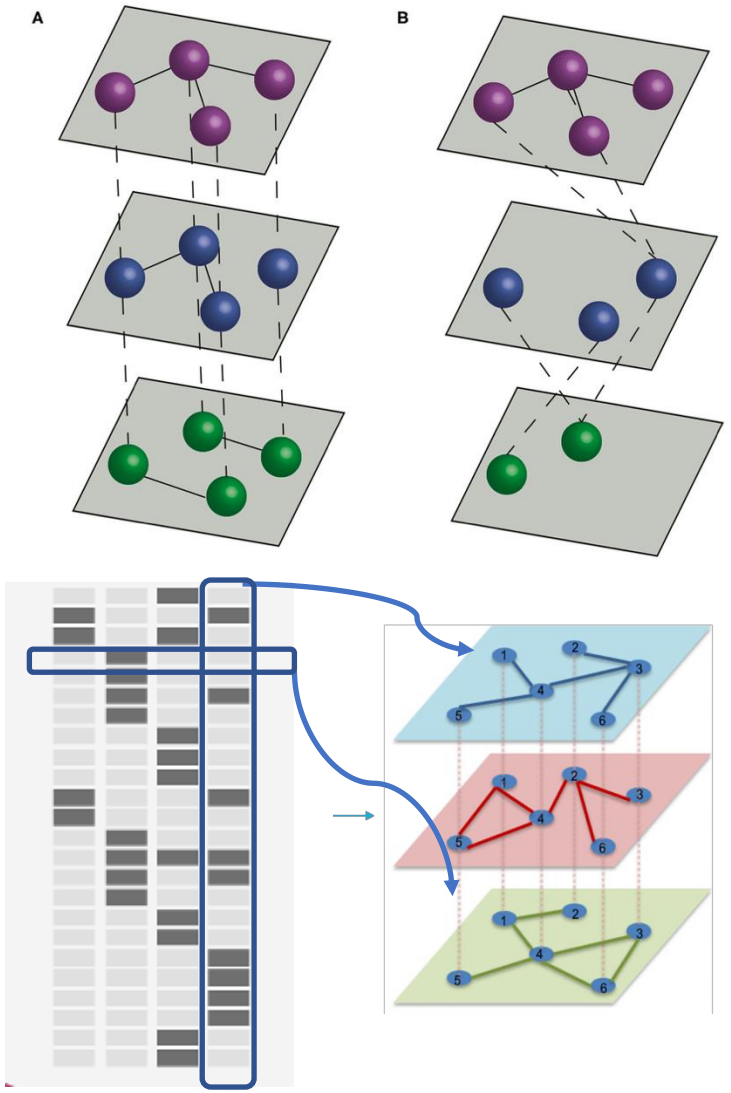
Mixed-modality Multi-graph:
Each colour is a distinct interaction pattern

$$\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \mathbf{A}^{(3)} \in \mathcal{R}^{P \times P}$$

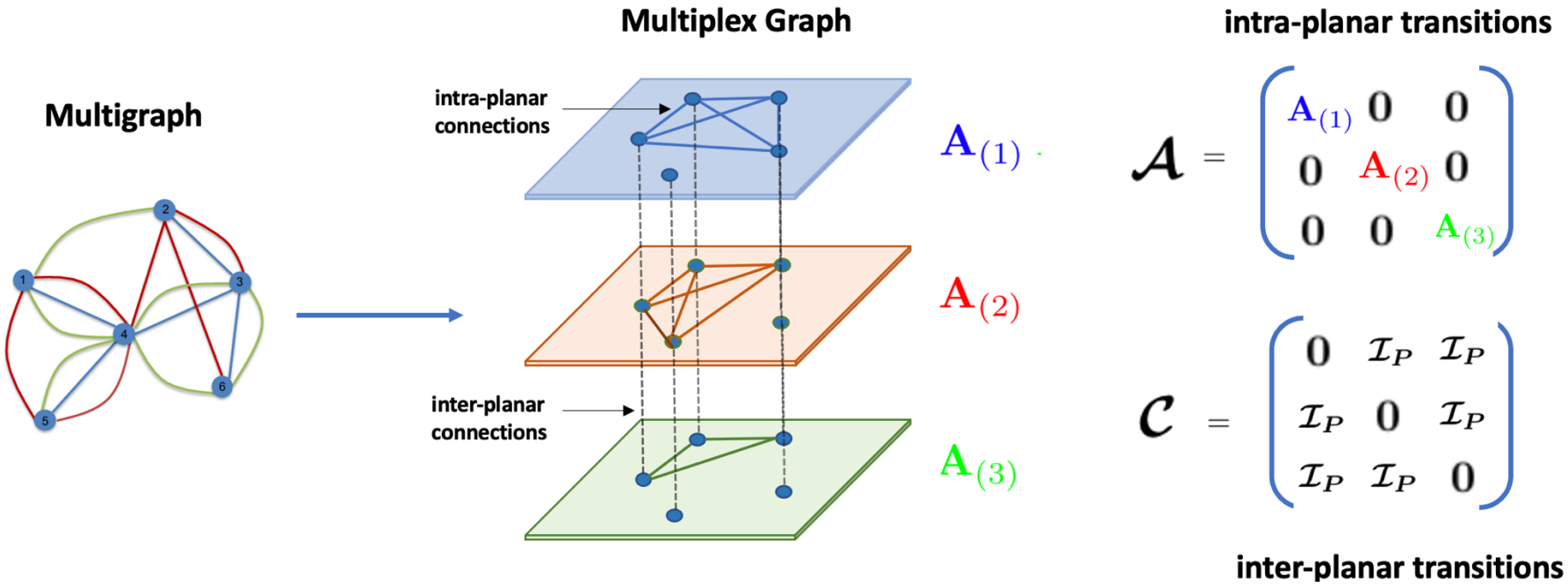
Multi-layer Networks – background

- Multilayer networks consist of nodes and edges, but the nodes exist in separate layers, representing different forms of interactions representing an aspect
  - e.g. different types of contacts, spatial locations, or points in time.
- Multilayer networks are of 2 types:
  - Multiplexed network
    - interlayer edges can only connect nodes that represent the same actor in different layers.
  - Interconnected network
    - In *interconnected networks,* interlayer edges can connect between different actors

Multiplexed network is more suitable for our formulation as each graph is all taken from one entity (e.g patient)



The Multiplex Framework



Motivation: Using intra- and inter "planar" connections to model different interaction patterns during multimodal fusion

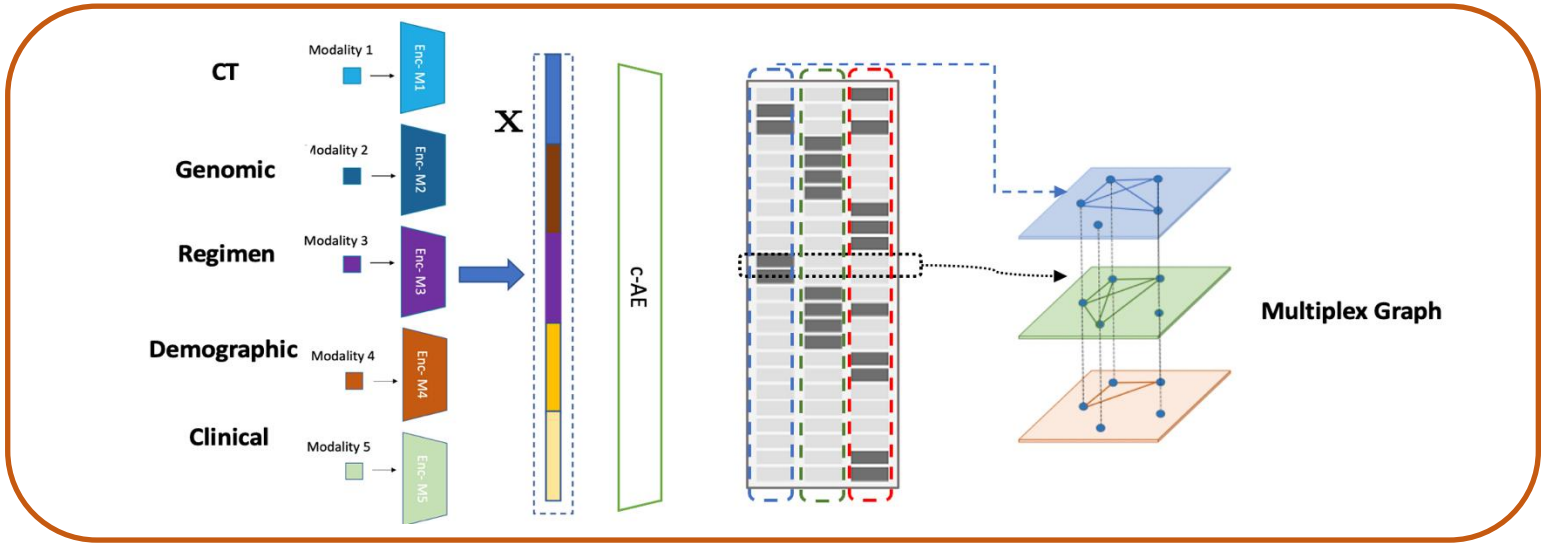
Used earlier for transportation networks or social networks

Multimodal Fusion via Multiplex Graphs

Multimodal Graph Representation Learning

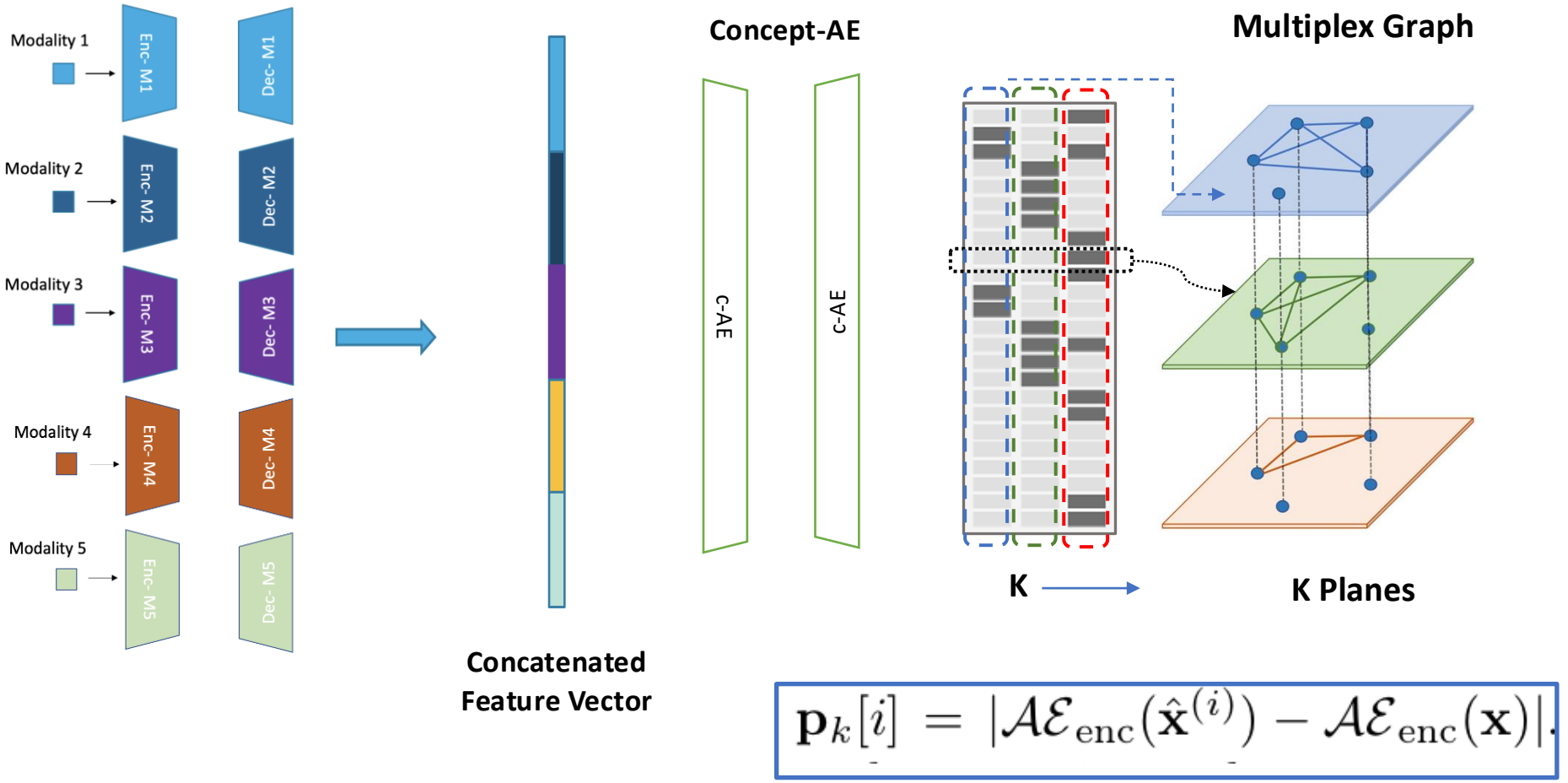
Outcome Prediction: Multiplex GNN

Multiplexed Graph Representation Learning



Outcome Prediction: Multiplex GNN

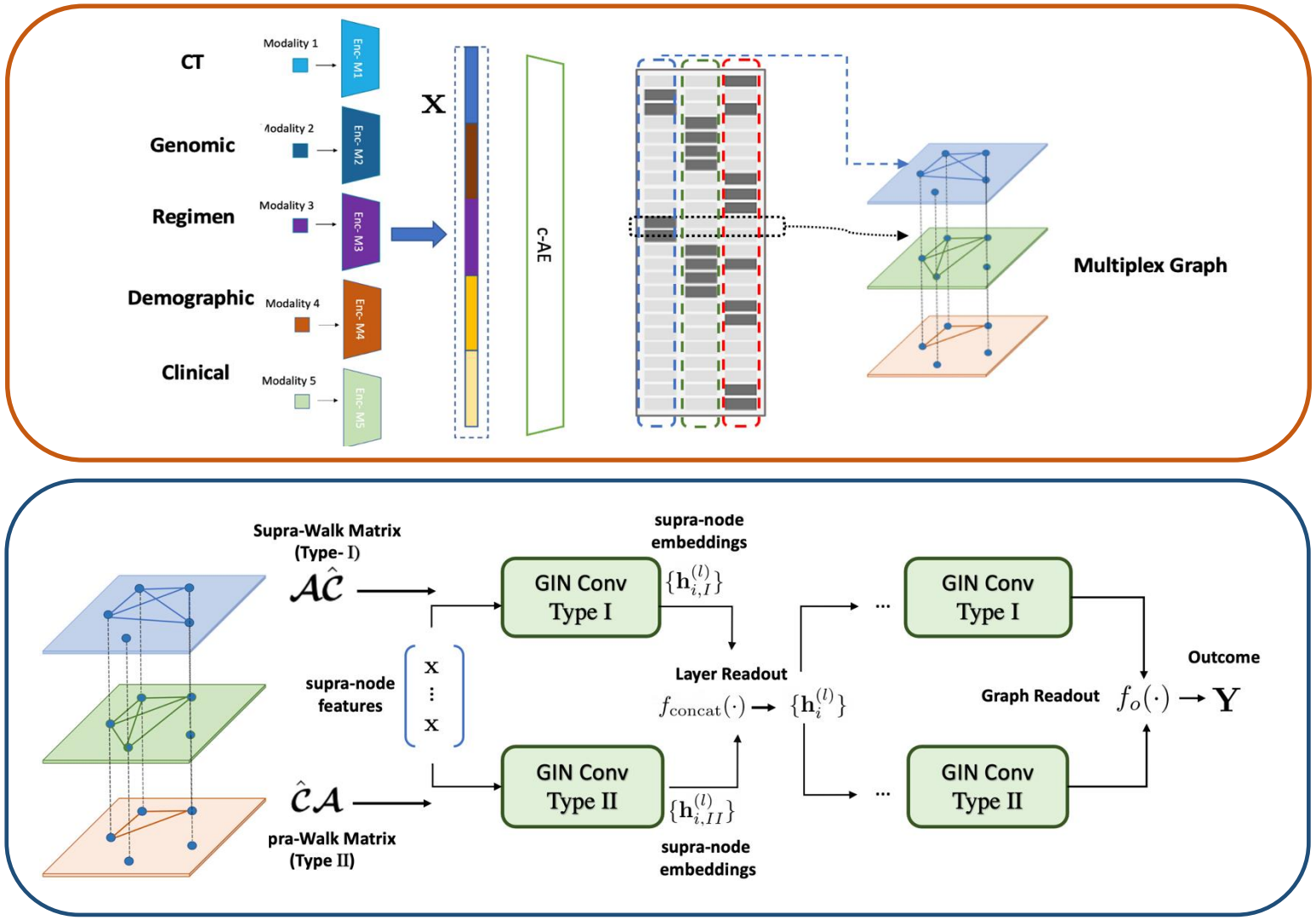
Multiplexed Graph Representation Learning



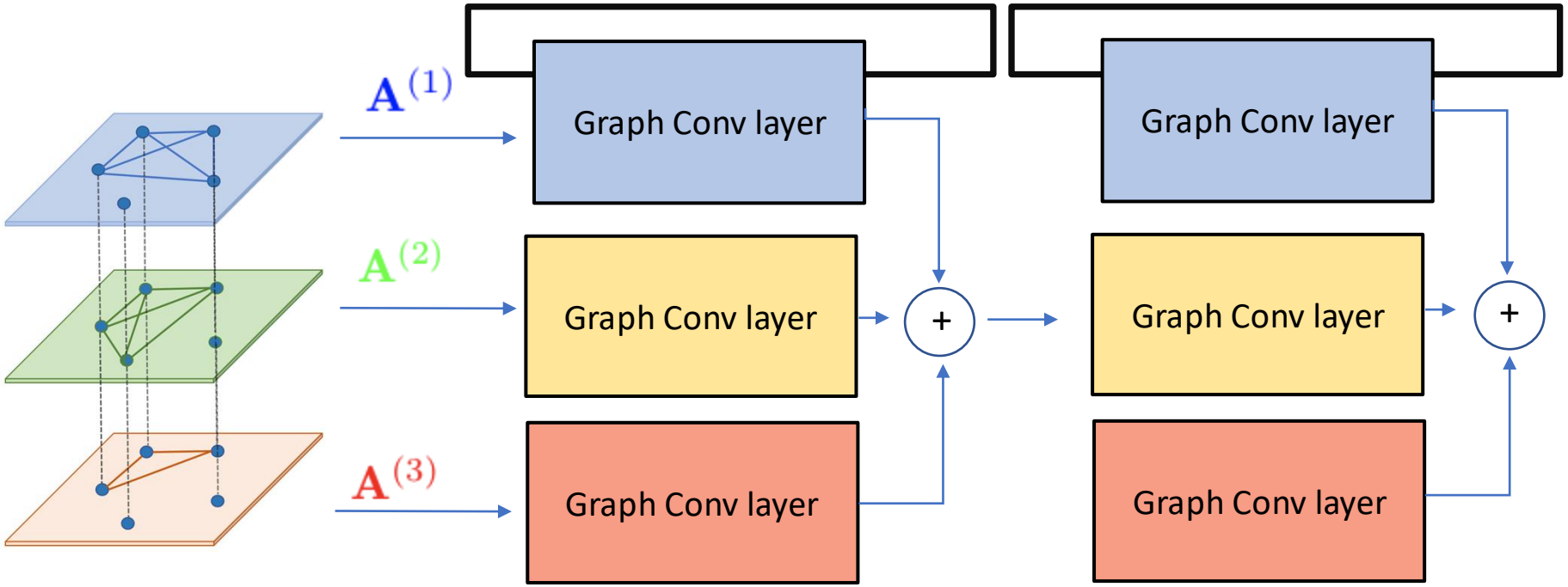
$$\mathbf{p}_k[i] = |\mathcal{AE}_{\text{enc}}(\hat{\mathbf{x}}^{(i)}) - \mathcal{AE}_{\text{enc}}(\mathbf{x})|.$$

Each latent dimension of the autoencoder captures an abstract aspect of the interaction pattern

Multiplexed Graph Neural Networks



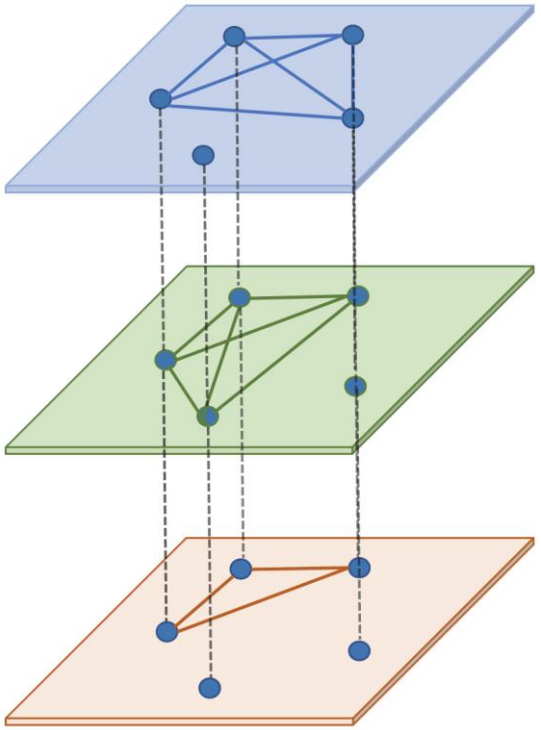
Prior Work on Multigraph Learning



Message Passing: $$\mathbf{H}_i^{(l+1)} = \sigma\Big[\sum_{k=1}^{K}\sum_{j\in N(i)}\frac{1}{c_{i,k}}\mathbf{H}_j^{(l)}\mathbf{W}_k^{(l)} + \mathbf{H}_i^{(l)}\mathbf{W}_0\Big]$$

- Separates message passing across edge-types and then aggregate
- May miss several higher order paths arising from cross modal connections

Defining Walks on the Multiplex



Intra-planar adjacency matrix

$$\mathcal{A} = \bigoplus_k \mathbf{A}^{(k)}$$

Inter-planar transition matrix

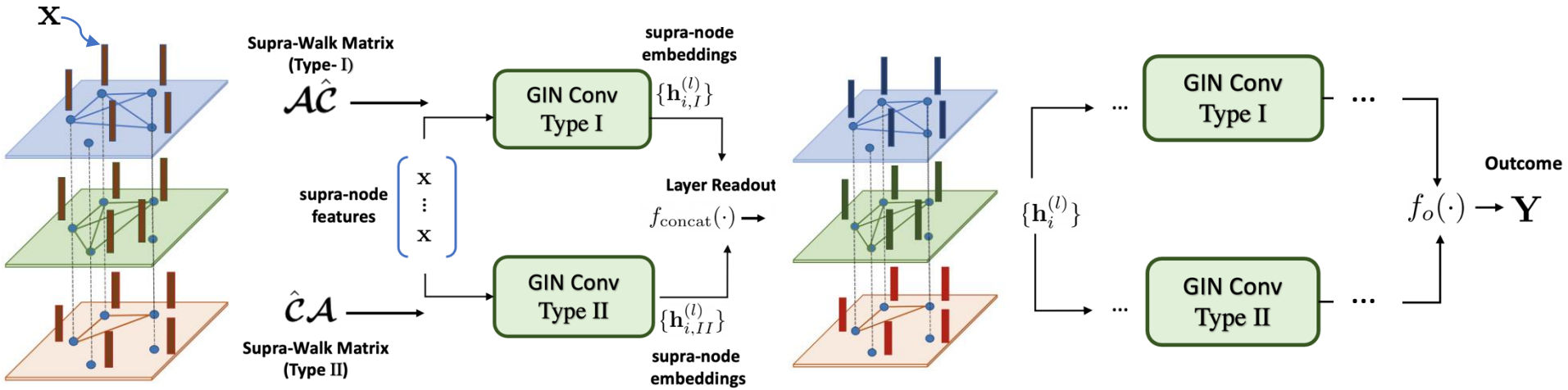
$$\mathcal{C} = [\mathbf{1}\mathbf{1}^T] \otimes \mathcal{I}_P - \mathcal{I}_{PK}$$

A walk consists of one of two types of steps:

I. First take intra-planar step, then allowed to transition $\mathcal{A}\hat{\mathcal{C}}$

II. First allowed to transition, then take intra planar step $\hat{\mathcal{C}}\mathcal{A}$

Multiplex Graph Neural Network



$$\mathbf{h}_{i,I}^{(l+1)} = \text{GINConv}\left(\{\mathbf{h}_j^{(l)}, j : [\mathcal{A}\hat{\mathcal{C}}][i,j] = 1\}\right) \quad \text{Type I Filtering}$$

$$\mathbf{h}_{i,II}^{(l+1)} = \text{GINConv}\left(\{\mathbf{h}_j^{(l)}, j : [\hat{\mathcal{C}}\mathcal{A}][i,j] = 1\}\right) \quad \text{Type II Filtering}$$

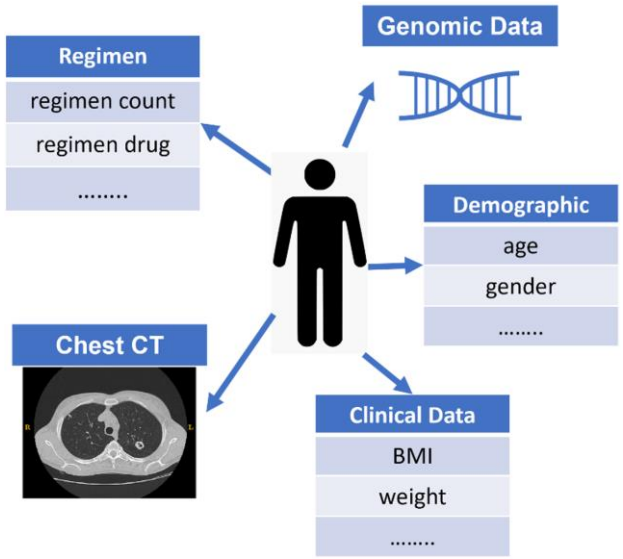
$$\mathbf{h}_i^{(l+1)} = f_{\text{concat}}(\mathbf{h}_{i,I}^{(l+1)}, \mathbf{h}_{i,II}^{(l+1)}) \quad \text{Layer-wise readout}$$

$$f_{\text{o}}(\{\mathbf{h}_i^{(L)}\}) = \mathbf{Y} \quad \text{Graph readout}$$

TB Dataset

Five modalities: [1 multiplex graph per patient]

- Imaging (CT annotations) – 2084 DenseNet features
- Genomic (SNP) - 4048 categorical features
- Demographic – 29 categorical, 1 continuous
- Clinical – 1726 categorical, 1 continuous
- Regimen – 233 categorical



Five Class Classification: [Graph level outcome] Cured, Failure, Still on treatment, Completed, Died

Baseline Comparisons

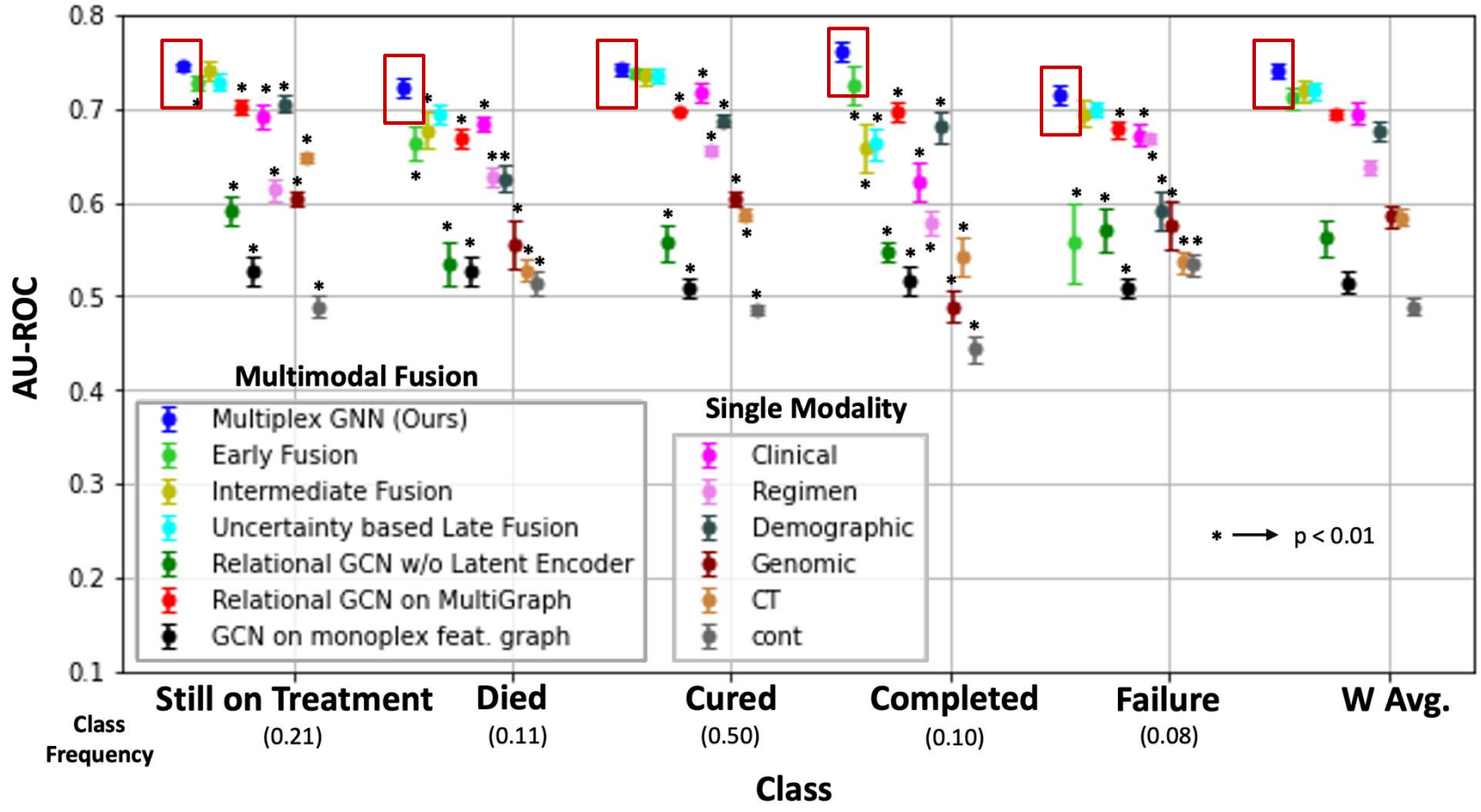
Fusion Baselines:

- No fusion /Individual modality prediction
- Early fusion
- Intermediate fusion
- Uncertainty based late fusion (Wang et. al. 2021)

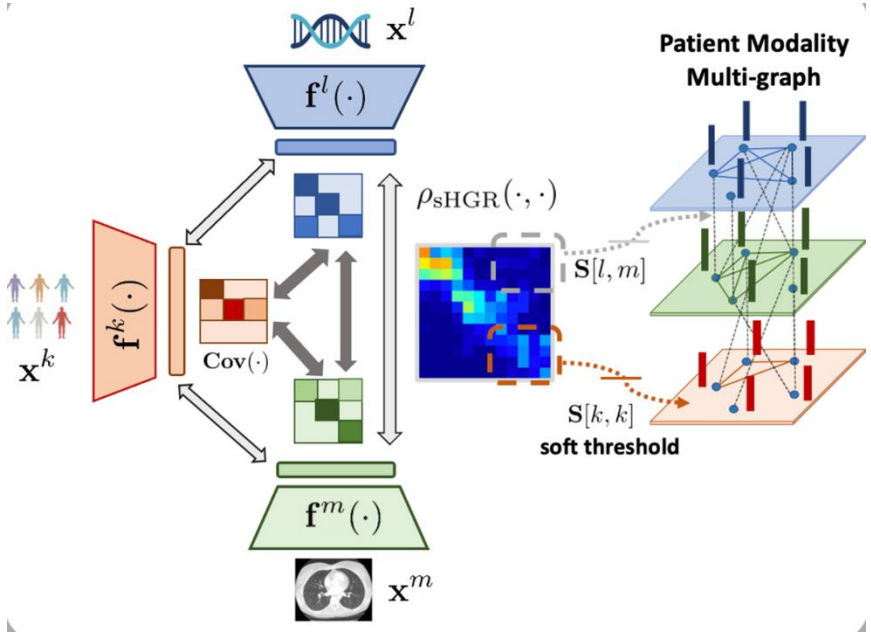
Ablations:

- Relational GCN on a Multiplexed Graph (Schlichtkrull et. al. ,2018)
- Relational GCN w/o Latent Encoder (Schlichtkrull et. al. ,2018)
- GCN on monoplex feature graph (Kipf and Weling et. al., 2016)

Results


AU-ROC
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Multimodal Fusion
Multiplex GNN (Ours)
Early Fusion
Intermediate Fusion
Uncertainty based Late Fusion
Relational GCN w/o Latent Encoder
Relational GCN on MultiGraph
GCN on monoplex feat. graph
Single Modality
Clinical
Regimen
Demographic
Genomic
CT
cont
* ⟶ p < 0.01
Still on Treatment
Died
Cured
Completed
Failure
W Avg.
Class Frequency
(0.21)
(0.11)
(0.50)
(0.10)
(0.08)
Class

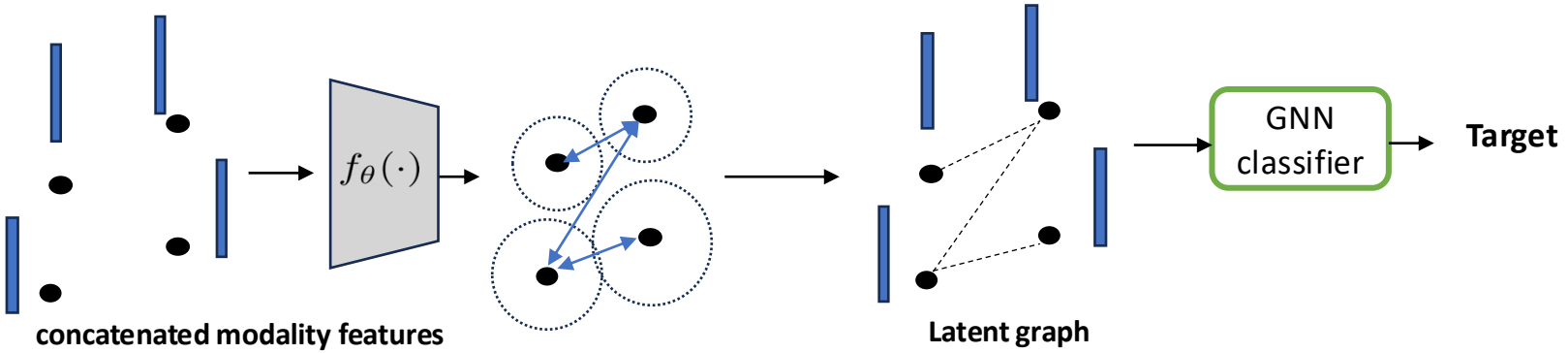
Latent Multi-Graph Learning: MaxCorr



Outcome Prediction:
Multi-Graph Neural Network

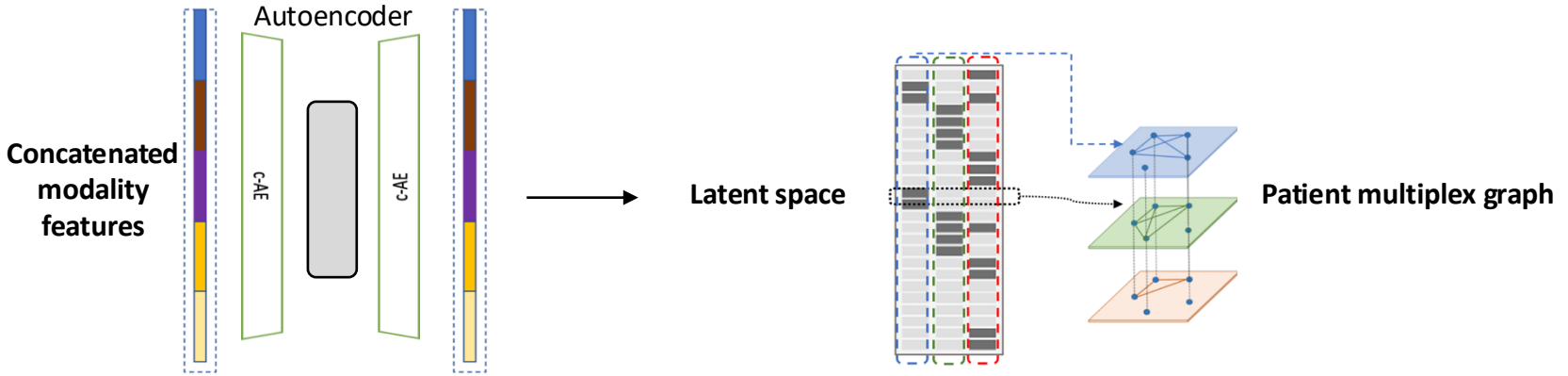
Graph Structure Learning for Fusion

Supervised Latent graph learning (Cosmo et al, 2020)



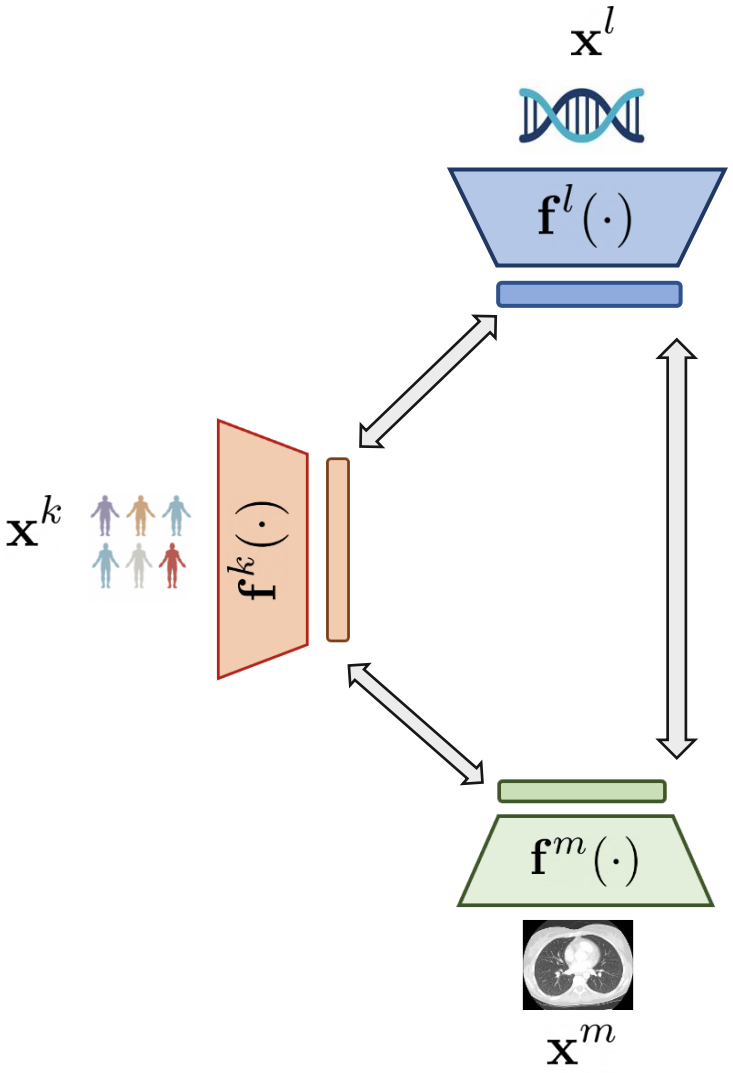
Limitation: Ad-hoc blending of modality representations

Unsupervised Multiplexed Graph Construction (Dsouza et al 2022)



Limitation: Not differentiable, i.e. cannot be trained end-to-end with task supervision

Maximal Correlations



Hirschfield Gebelin Renyi (HGR)

The transformations f an g are non-linear and most "informative" ones, in the view of information theory, as f(X) carries the maximum amount of the information towards Y and vice versa

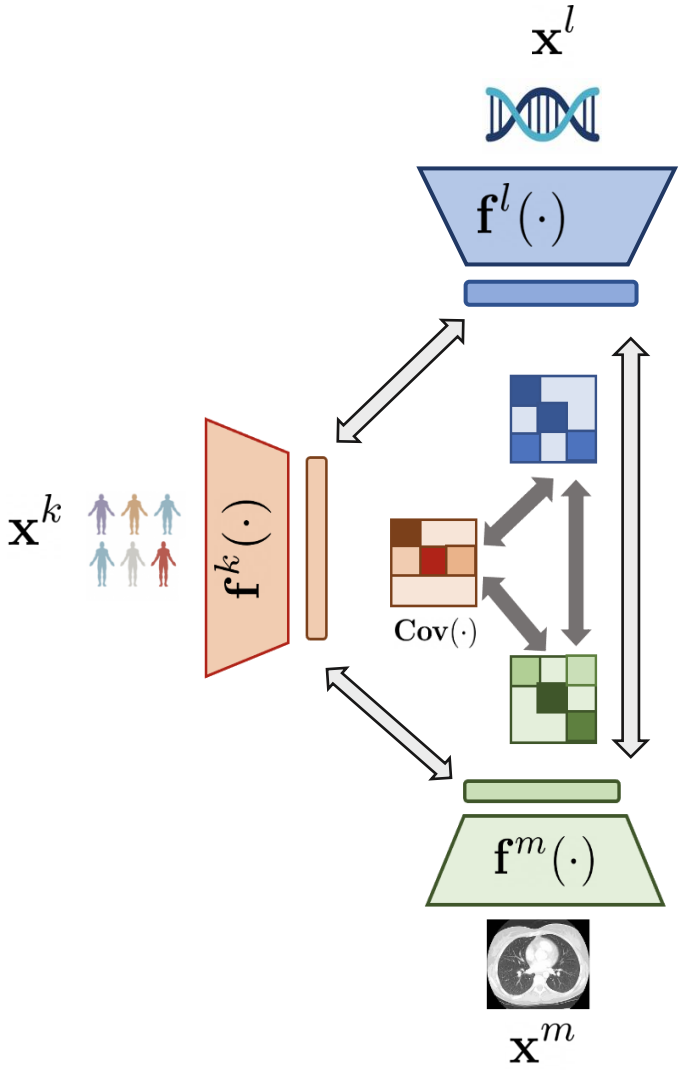
$$\sup_{\mathcal{C}_E,\mathcal{C}_{\text{Cov}}} \rho_{\text{HGR}}(\mathbf{x}^l,\mathbf{x}^m) = \sup_{\mathcal{C}_E,\mathcal{C}_{\text{Cov}}} \mathbb{E}\Big[[\mathbf{f}^l(\mathbf{x}^l)]^T\mathbf{f}^m(\mathbf{x}^m)\Big]$$

$$\mathcal{C}_E : \{\mathbb{E}[\mathbf{f}^l(\cdot)] = \mathbb{E}[\mathbf{f}^m(\cdot)] = \mathbf{0}\}$$
$$\mathcal{C}_{\text{Cov}} : \{\mathbf{Cov}(\mathbf{f}^l(\mathbf{x}^l)) = \mathbf{Cov}(\mathbf{f}^m(\mathbf{x}^m)) = \mathcal{I}_{D_p}\}$$

Challenges:

- Whitening constraints require matrix inversion
- If the modalities are weakly correlated then HGR may not work well for downstream discriminative tasks

Soft HGR loss

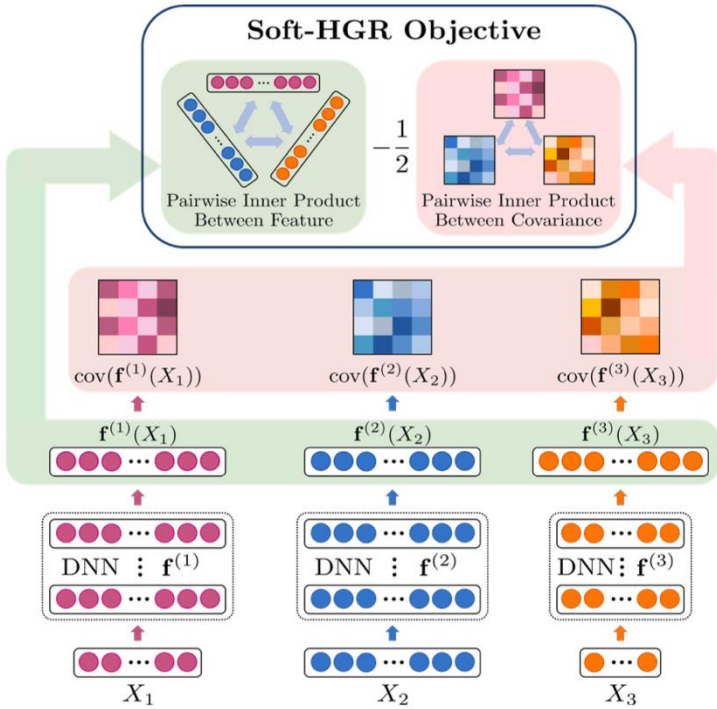


Soft Hirschfield Gebelin Renyi (sHGR)

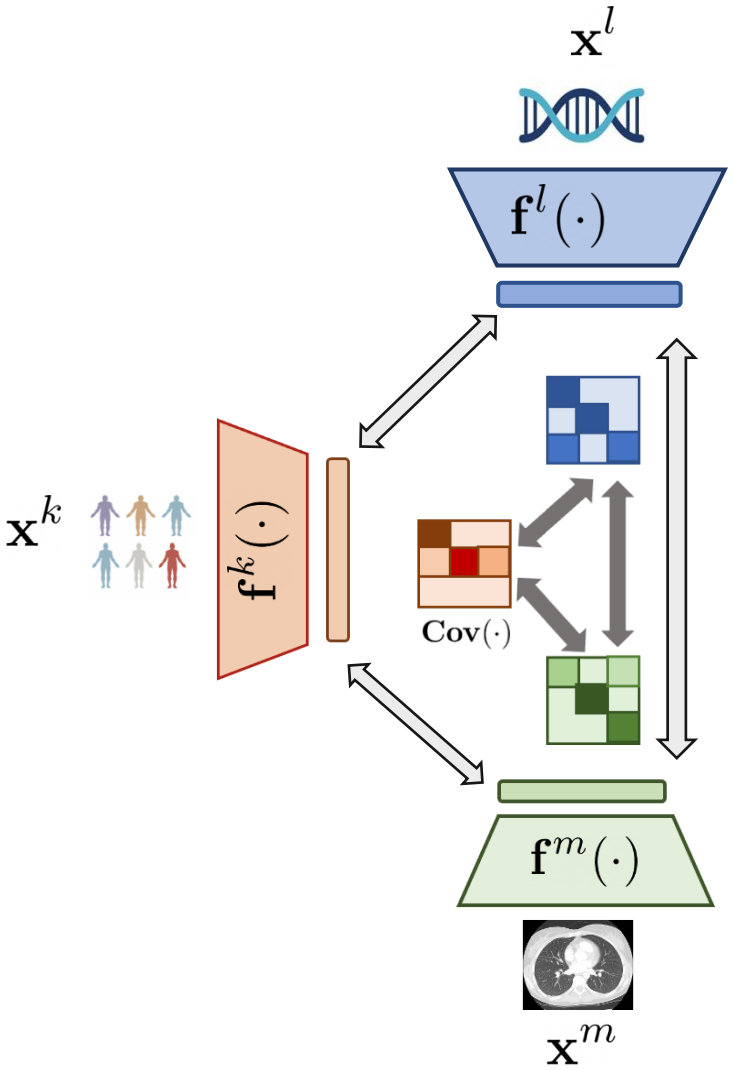
$$\mathcal{L}_{\text{sHGR}} = -\frac{1}{N_z}\mathbb{E}\Big[\mathbf{f}^l(\mathbf{x}^l)^T\mathbf{f}^m(\mathbf{x}^m)\Big] + \frac{1}{2N_z}\,\mathrm{Tr}\,\Big[\mathbf{Cov}[\mathbf{f}^l(\mathbf{x}^l)]\mathbf{Cov}[\mathbf{f}^m(\mathbf{x}^m)]\Big]$$
$$l \neq m$$

low-rank approximation, to approach the HGR problem

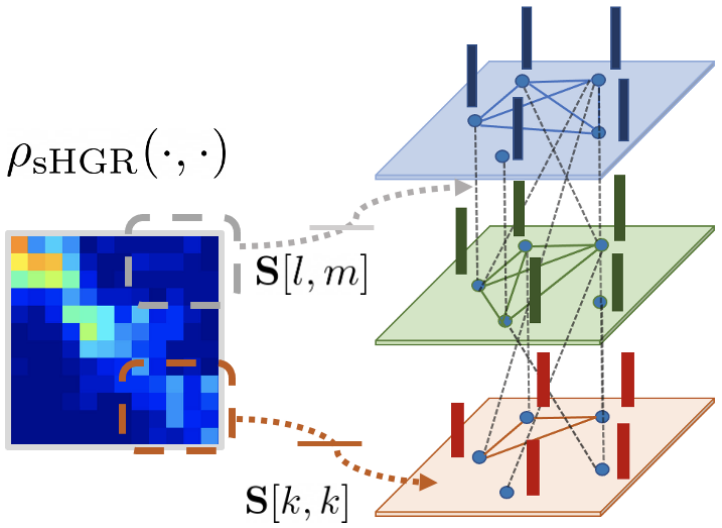
As in Deep CCA, f(x) estimated through non-linear parametric neural networks.



Adaptive Sparse Thresholding



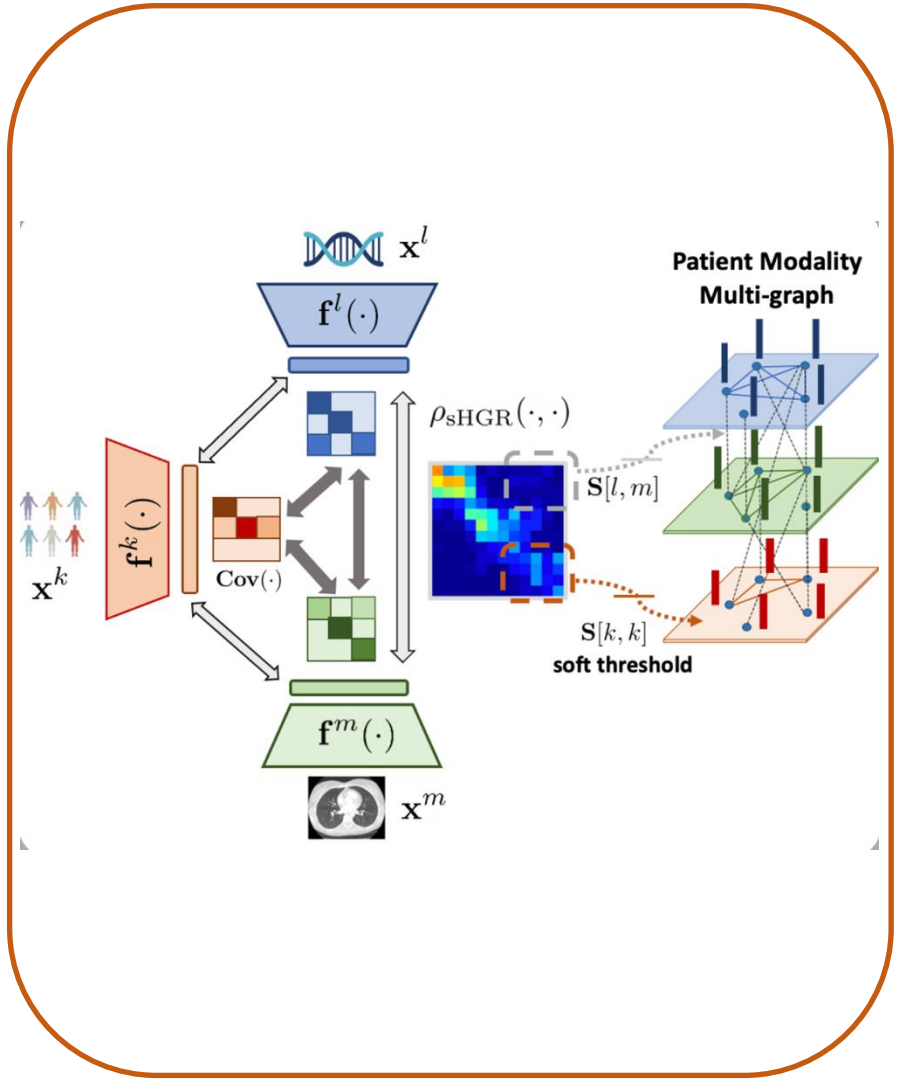
$$\mathbf{C}_{(l,m)}[i,j] = \mathrm{ReLU}(\tilde{\rho}_{\mathrm{sHGR}}(\mathbf{x}_i^l, \mathbf{x}_j^m) - \tilde{\mathbf{S}}[l,m])$$
$$\mathbf{A}_k[i,j] = \mathrm{ReLU}(\tilde{\rho}_{\mathrm{sHGR}}(\mathbf{x}_i^k, \mathbf{x}_j^k) - \tilde{\mathbf{S}}[k,k])$$

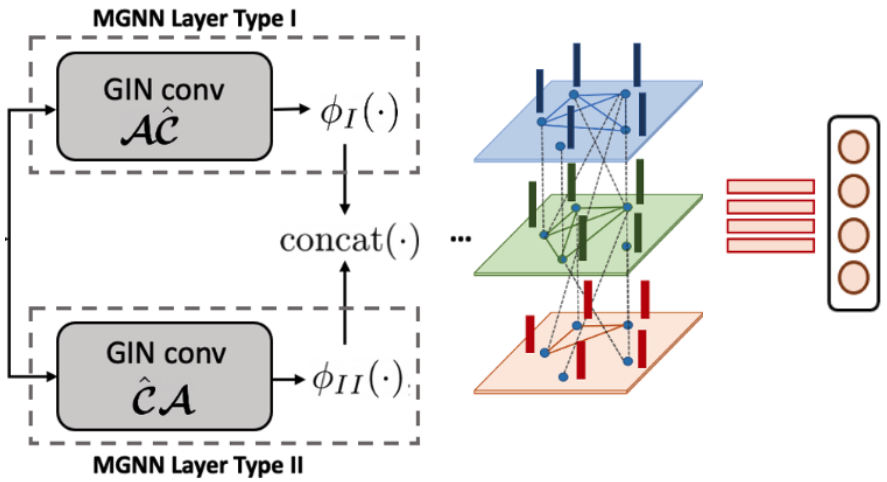


Patient-Modality Multi-Graph

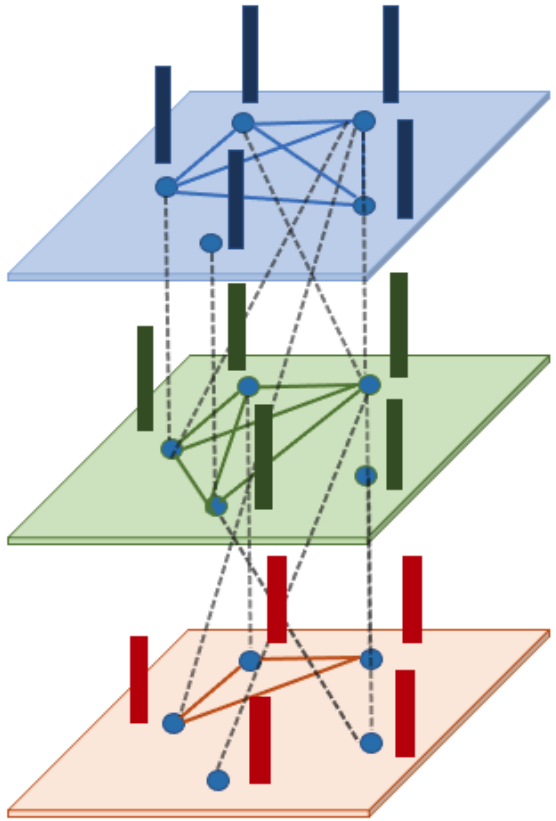
Nodes --- > Patients
Planes --- > Modalities
Edge weights --- > sHGR Correlations

MGNN for Outcome Prediction





Walks on the Multi-Graph



Intra-planar adjacency matrix

$$\mathcal{A} = \bigoplus_{k} \mathbf{A}^{(k)}$$

Inter-planar transition matrix

$$\hat{\boldsymbol{\mathcal{C}}}$$

$$\hat{\boldsymbol{\mathcal{C}}}[lP:(l+1)P, mP:(m+1)P]$$

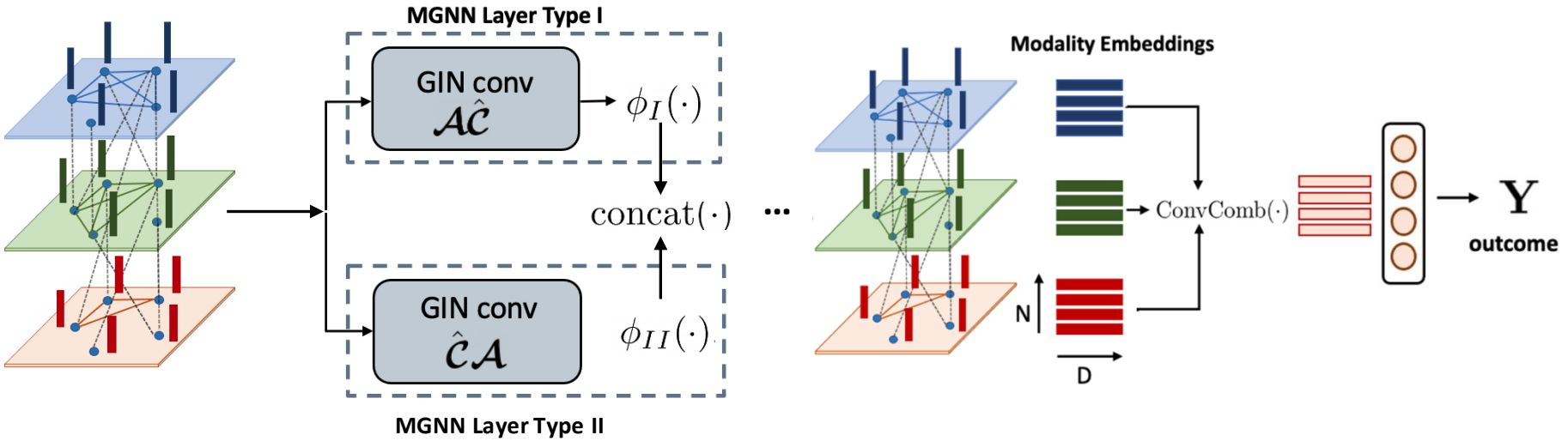
$$= \mathbf{C}_{(l,m)}\ \mathbb{1}(l \neq m) + \boldsymbol{\mathcal{I}}_{\boldsymbol{P}}\ \mathbb{1}(l = m)$$

A walk consists of one of two types of steps:

I. First take intra-planar step, then allowed to transition $\mathcal{A}\hat{\mathcal{C}}$

II. First allowed to transition, then take intra planar step $\hat{\mathcal{C}}\mathcal{A}$

Multi-Graph Neural Network



$$\mathbf{h}_{s_i,I}^{(d+1)} = \boldsymbol{\phi}_I^{(d)}\left((1+\epsilon)\mathbf{h}_{s_i}^{(d)} + \text{wmean}\left[\mathbf{h}_{s_j}^{(d)}, \boldsymbol{\mathcal{A}\hat{\mathcal{C}}}[s_i,s_j] \;\; ; \;\; s_j \in \mathcal{N}_{\boldsymbol{\mathcal{A}\hat{\mathcal{C}}}}(s_i)\right]\right)$$ **MGNN Type I Filtering**

$$\mathbf{h}_{s_i,II}^{(d+1)} = \boldsymbol{\phi}_{II}^{(d)}\left((1+\epsilon)\mathbf{h}_{s_i}^{(d)} + \text{wmean}\left[\mathbf{h}_{s_j}^{(d)}, \boldsymbol{\hat{\mathcal{C}}\mathcal{A}}[s_i,s_j] \;\; ; \;\; s_j \in \mathcal{N}_{\boldsymbol{\hat{\mathcal{C}}\mathcal{A}}}(s_i)\right]\right)$$ **MGNN Type II Filtering**

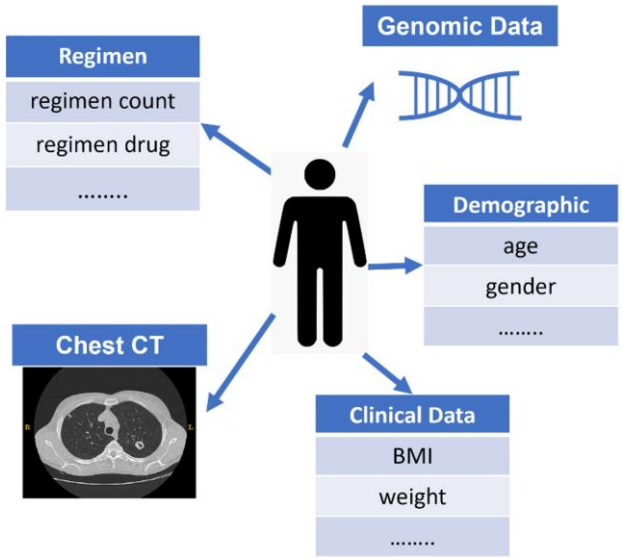
$$\mathbf{h}_{s_i}^{(d+1)} = \text{concat}(\mathbf{h}_{s_i,I}^{(d+1)}, \mathbf{h}_{s_i,II}^{(d+1)})$$ **Layer-wise readout**

$$\mathbf{g}_o(\{\mathbf{h}_{s_i}^{(L)}\}_{s_i \leftrightarrow i}) = \hat{\mathbf{Y}}_i$$ **Classifier**

TB Dataset

Five modalities: [Patient –modality multi-graph]

- Imaging (CT annotations) – 2084 DenseNet features
- Genomic (SNP) - 4048 categorical features
- Demographic – 29 categorical, 1 continuous
- Clinical – 1726 categorical, 1 continuous
- Regimen – 233 categorical



Five Class Classification: [Node level outcome] Cured, Failure, Still on treatment, Completed, Died

Baseline Comparisons

No Fusion Baselines: Individual modality prediction

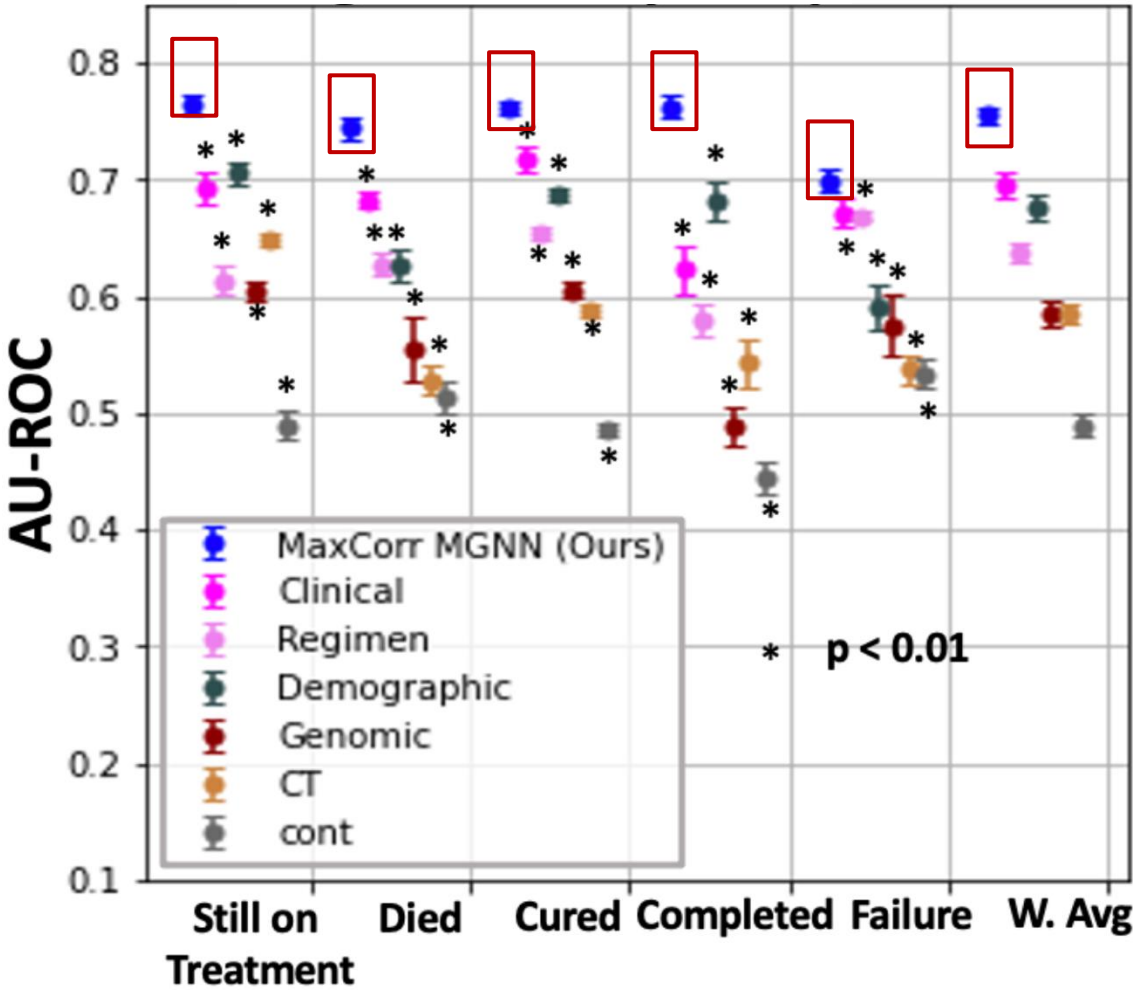
Fusion Baselines:

- Early fusion
- Intermediate fusion (Dsouza et al 2022)
- Uncertainty based late fusion (Wang et. al. 2021)
- Latent Graph Learning for fusion (Cosmo et. al 2020)

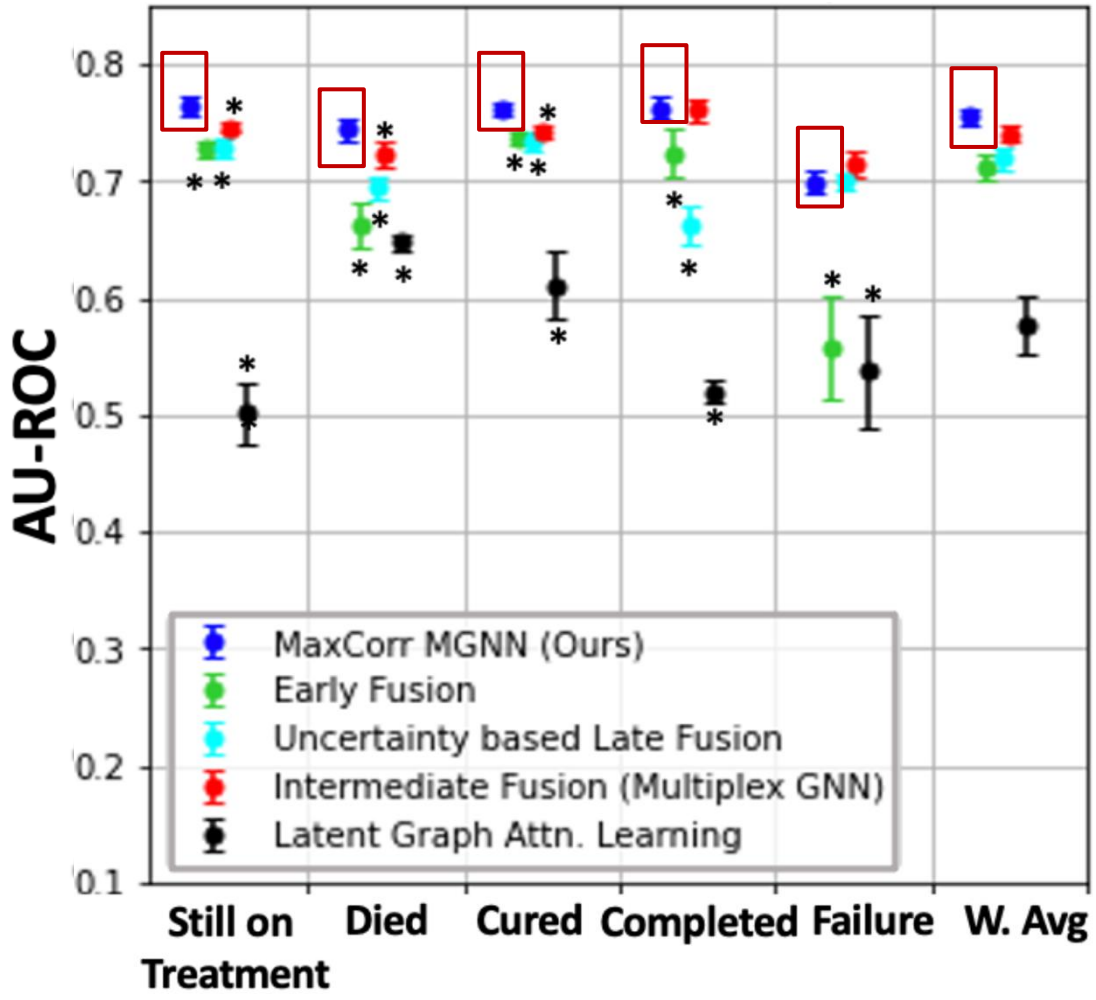
Ablations:

- sHGR+ MLP (Wang et al. 2019)
- MaxCorrMGNN w/o sHGR
- Decoupled MaxCorrMGNN

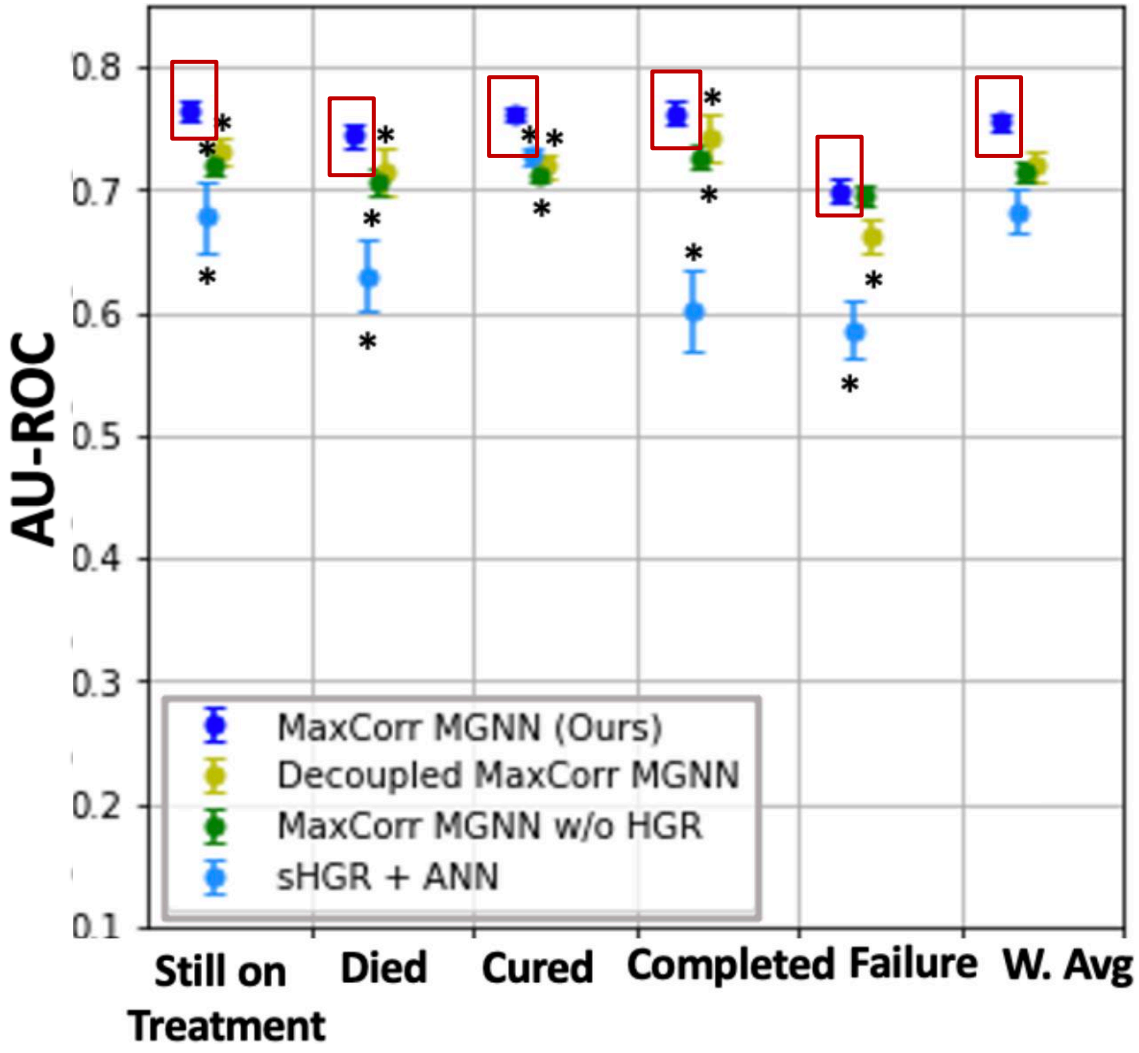
Results – Single Modality Comparisons



Results – Fusion Baselines



Results – Ablations

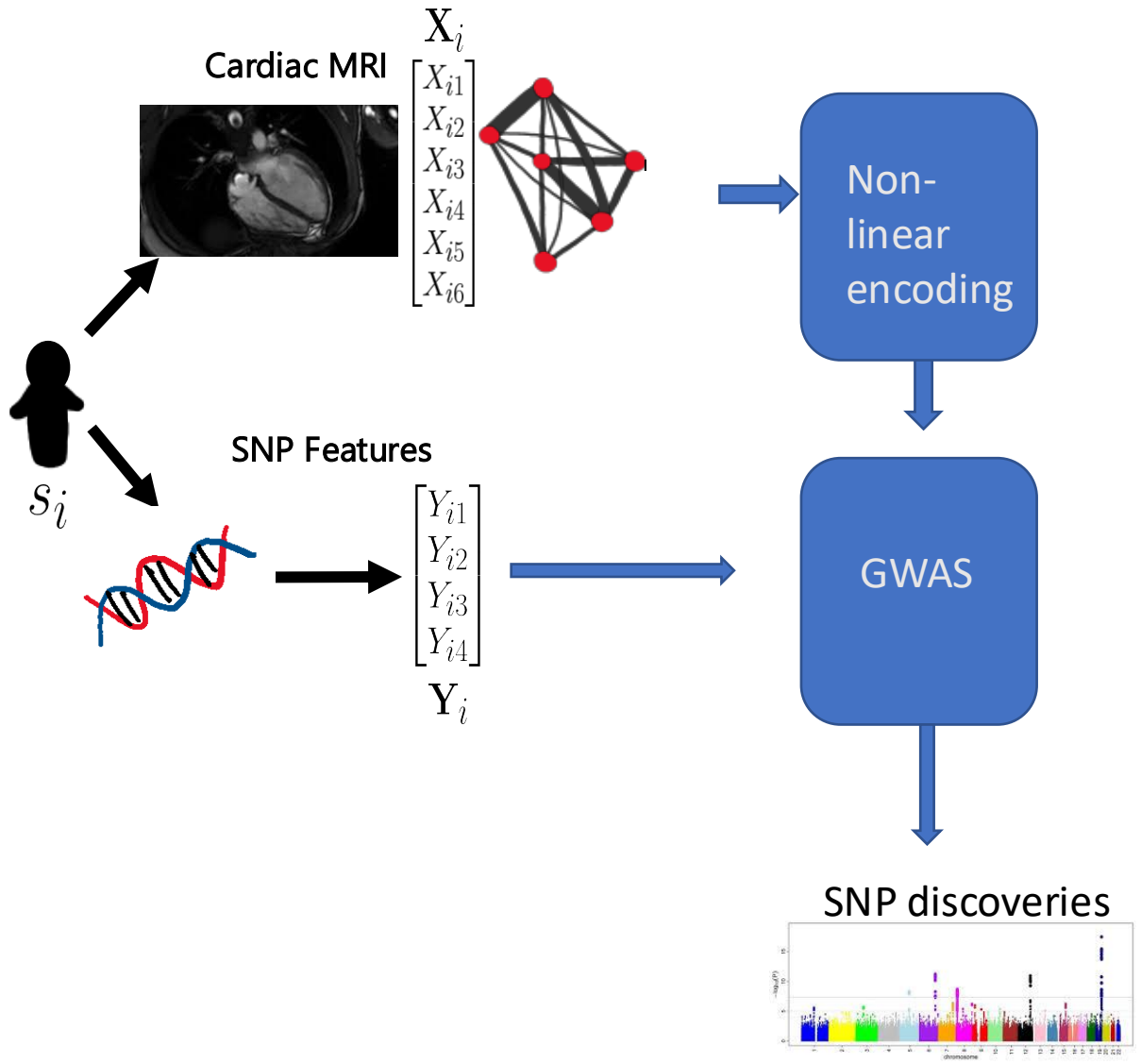


MaxCorr GNN: A End-to-End Generalized Multimodal Fusion Framework

- Learns task informed patient-modality multi-graph representations automatically from unstructured modality data
- Multilayered graph representation keeping the modalities separate
- Formalism for maximal correlation (Hirschfield, Gebelin, Rènyi (HGR)) to build the graph representation
- Task-specific inference through graph neural networks
- Loss function for end-to-end learning of both graph connectivity and inference.

Can multimodal fusion lead to discovery?

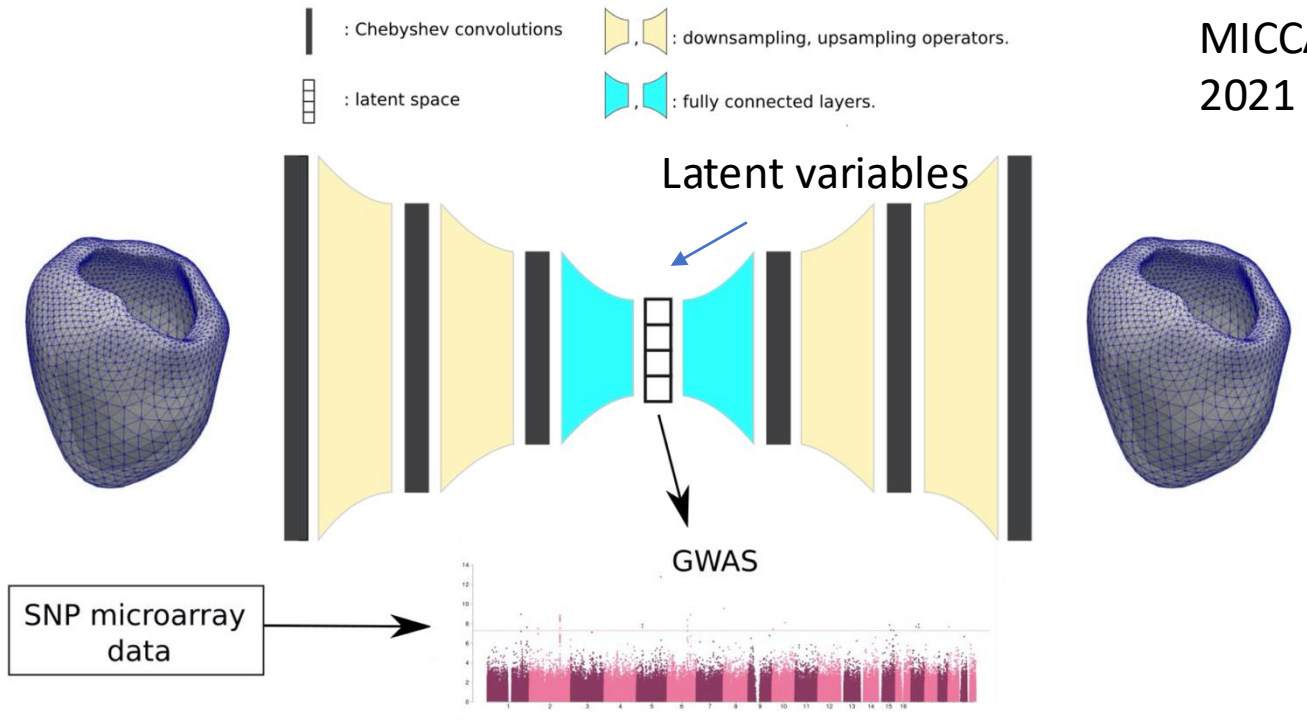
- Illustrated in the context of cardiovascular diseases
  - Non-linear correlation between modalities
  - Projections for one of the modalities but original features retained for the other modality
  - Allowing for per feature correlation could lead to novel discoveries through GWAS
  - GWAS searches across their entire genome for SNP and spots consistent differences.
  - GWAS done so far on handcrafted cardiac structure features (volume, mass)
  - Shape features encoded in a learnt latent space can provide a more informative imaging phenotype
  - What non-linear encoding could lead to non-spurious association discovery in GWAS?



Joint convolutional mesh autoencoder

- LV shape represented by a 3D mesh treated as a graph
- Eigenvalue decomposition of the Laplacian of the graph gives the basis for doing the graph convolutions

In MICCAI 2021



$$L = D - A, \qquad \tilde{L} = U \Lambda U^T$$

$$x * y = U((U^T x) \odot (U^T y))$$

Since this is computationally intensive, we apply a Chebyshev kernel

$$g_\theta(L) = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L}), \quad \tilde{L} = 2L/\lambda_{max} - I_n$$

Spectral convolution

$$y_j = \sum_{i=1}^{F_{in}} g_{\theta_{i,j}}(L) x_i \in \mathbb{R}^n$$

Added KL Divergence term to increase statistical independence of latent representations
Probability distribution on latent variables

	Input	Output
ChebConv	2677×3	2677×16
DS	2677×16	670×16
ChebConv	670×16	670×16
DS	670×16	168×16
ChebConv	168×16	168×16
DS	168×16	56×16
ChebConv	56×16	56×32
DS	56×32	28×32
ChebConv	28×32	28×32
FC	28×32	8×1

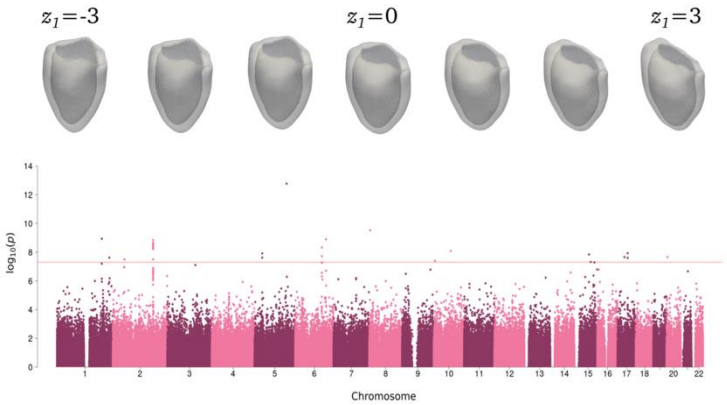
Results

Published in MICCAI 2021

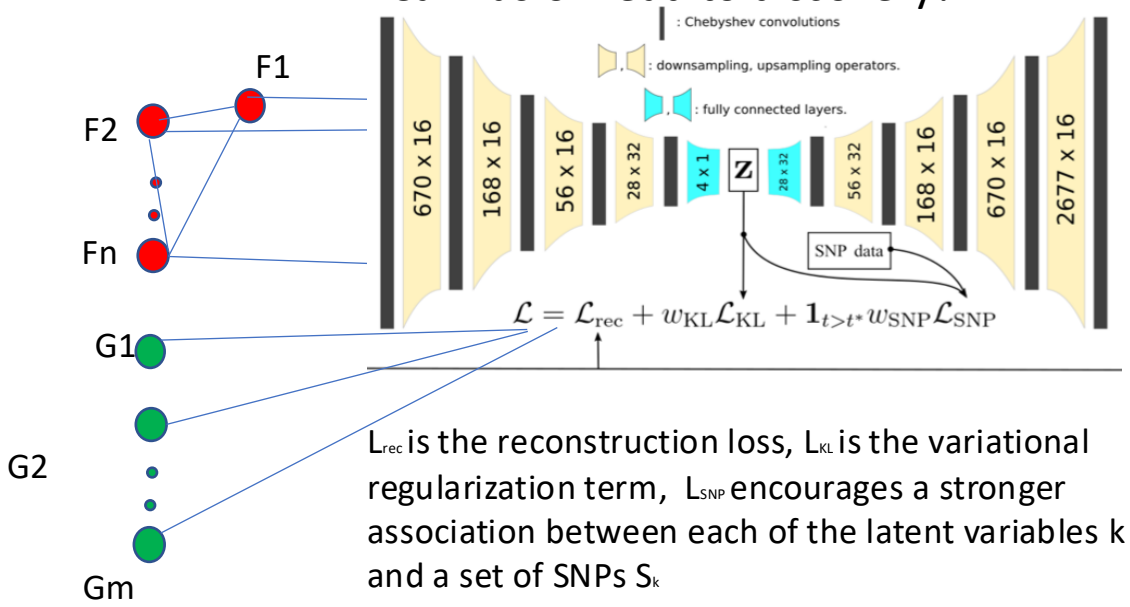
Dataset:

- UK Biobank
- 40,000 cardiac meshes
- SNP for all the patients
- GWAS across 29103 patients

Chromosome	Gene	P-value
5	ARHGAP26	10^{-13}
8	CSMD1	10^{-10}
6	PLN	10^{-9}
2	TTN	10^{-9}
6	PLN	10^{-8}



Can fusion lead to discovery?



L_{rec} is the reconstruction loss, L_{KL} is the variational regularization term, L_{SNP} encourages a stronger association between each of the latent variables k and a set of SNPs S_k

PLN gene to LV sphericity

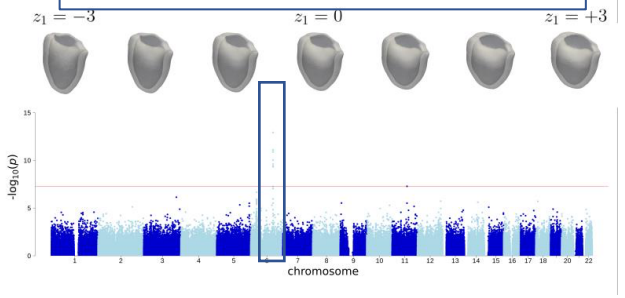


Fig. 2. Effect of latent variable z_1 (before re-training) and corresponding Manhattan plot.

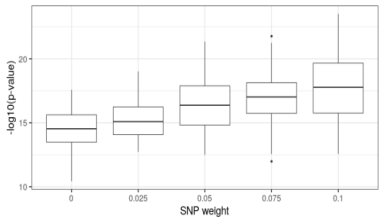
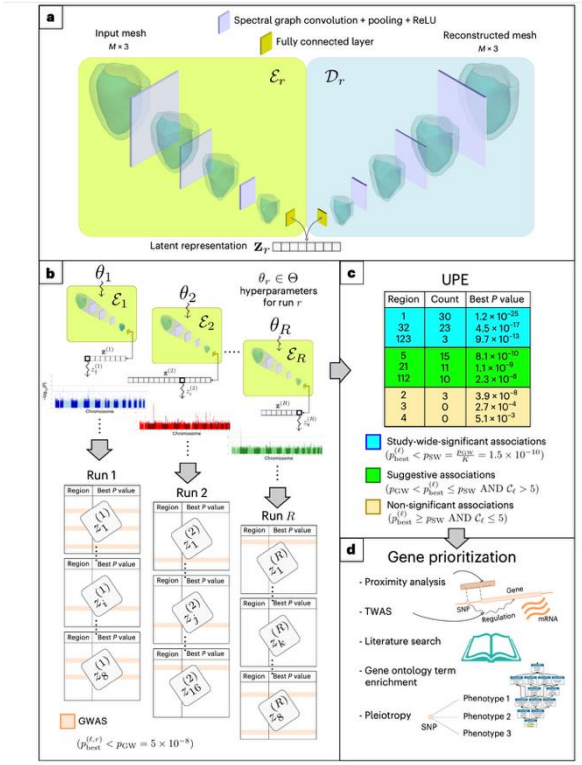
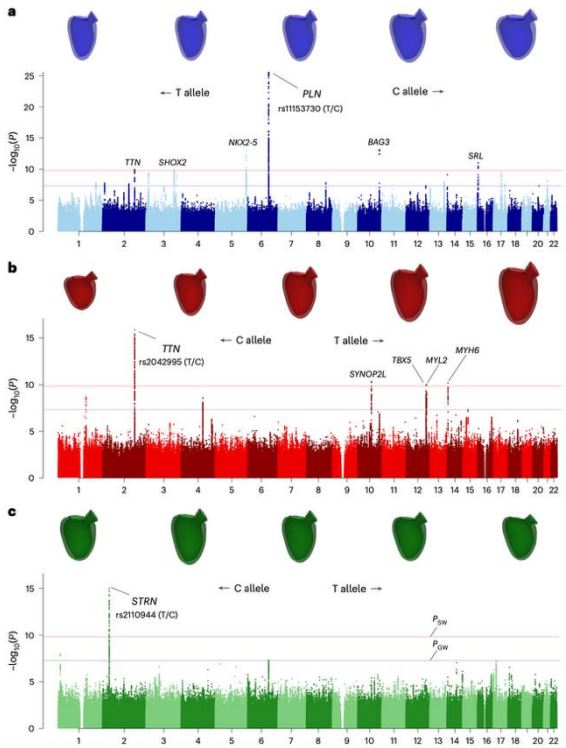


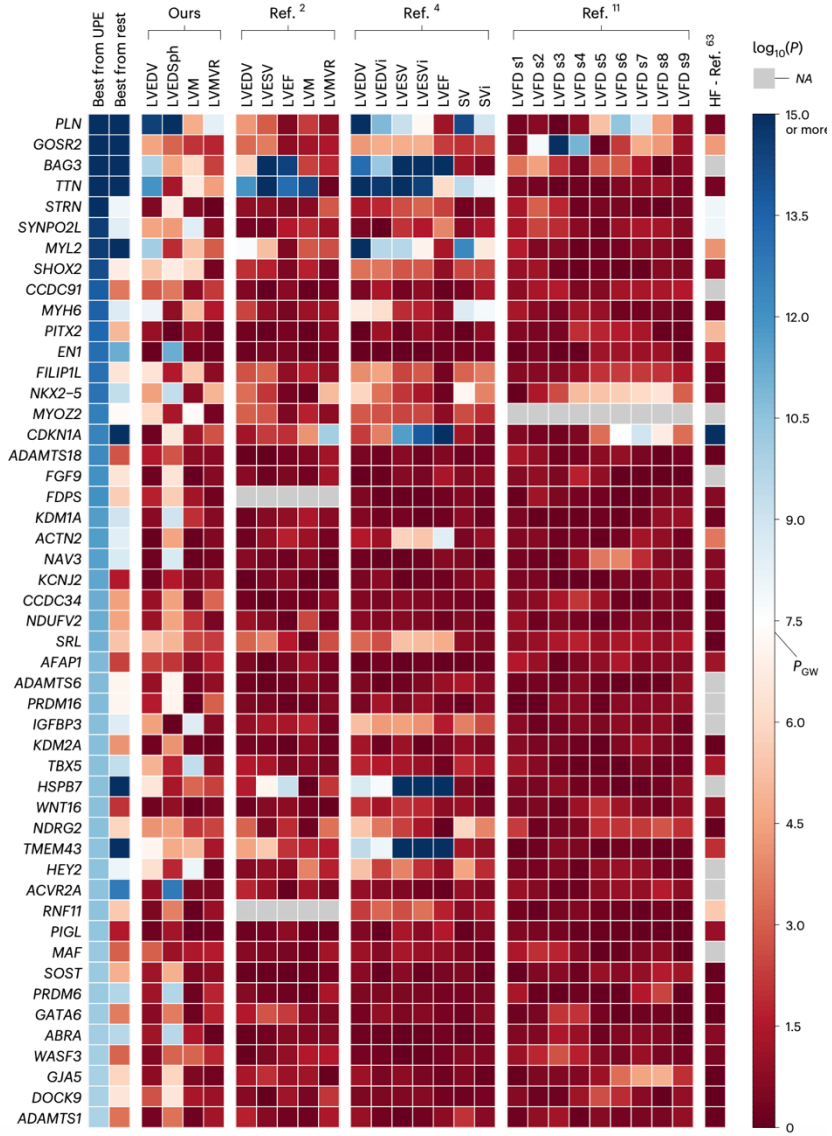
Fig. 3. Distribution of *p*-values for the z_1-rs11153730 associations after re-training, for experiments with different values of w_{SNP} with w_{KL} = 0.1. Each box contains 60± 10 experiments.

Correlation between z1 and LV sphericity

Gene discovery through latent shape associations







Loci with study-wide significance	Previously found	Handcrafted shape features	Shape PCA-based	Suggested needing verification
49	9	8	12	24

Multimodal fusion – Our Learning

- The key is to model the features, and their correlations
- The methods vary depending on the modalities:
  - Co-occur in space or time
    - Clinical and Imaging data in TB dataset
  - Provide complementary information
    - E.g. PET-MRI fusion
  - Provide contradictory information
  - Are mutually exclusive
  - Are mutually reinforcing
    - Majority of cases
  - Uncorrelated
  - Correlated
  - Confirmatory so that additional modalities don't make the sense
  - Combining them leads to new discoveries
  - Error-prone themselves
  - Have missing data
  - Have spurious data
- So the expected results from multimodal fusion could lead to:
  - Higher, lower, or about the same accuracy as individual modalities

- Metrics for multimodal fusion
  - Classification accuracy
  - Discovery potential
  - Additional insight about the findings in individual modalities
  - Conclusions about which of the cases of modality fusion is best suited for this problem