SUPREME: multiomics data integration using graph convolutional networks

Ziynet Nesibe Kesimoglu[1] and Serdar Bozdag[1,2,3,*]

BIBS Seminar
24/04/11
Hanbyul Song

서울대학교 통계학과
생물정보통계연구실
BIBS

Published online 28 June 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 2 **1**
https://doi.org/10.1093/nargab/lqad063

SUPREME: multiomics data integration using graph convolutional networks

Ziynet Nesibe Kesimoglu[1] **and Serdar Bozdag**[1,2,3,*]

[1]Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA, [2]Department of Mathematics, University of North Texas, Denton, TX, USA and [3]BioDiscovery Institute, University of North Texas, Denton, TX, USA

Received January 19, 2023; Revised May 08, 2023; Editorial Decision June 01, 2023; Accepted June 07, 2023

Contents

Introduction

Introduction:

- Cancer is one of the deadliest diseases for which cancer-causing agents such as oncogenes, mutations, and gene regulatory associations have not been fully understood.

- Patients show different characteristics in terms of the progression of disease and response to treatment.

- Large-scale datasets like **The Cancer Genome Atlas (TCGA)** and the **Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)** provide rich multiomics information, from genetic mutations to gene expression profiles.

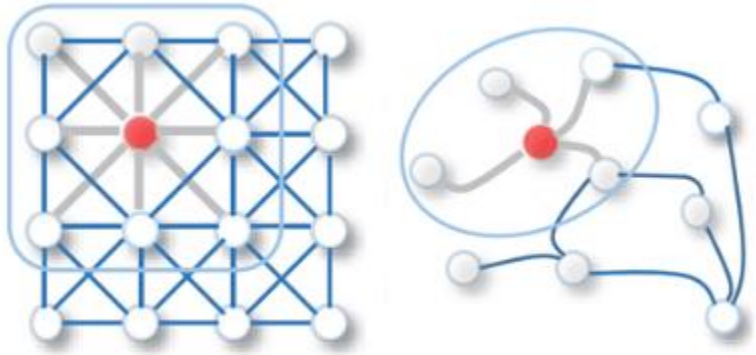
Introduction:

- Many cancer subtype prediction studies have relied solely on one type of biological data, which only captures a portion of the underlying biology.
- Existing computational tools struggle to integrate the comprehensive scope of multiomics data effectively.
- Previous methods like **iClusterPlus**, **Similarity Network Fusion (SNF)**, and **PINSPlus** have limitations in their integration approaches, focusing either on patient similarity networks or features without utilizing both.
- Limitations include the applicability to single networks, not leveraging multiple data modalities, and not utilizing node features comprehensively in network structures.

Introduction:

- **Understanding GNNs and Their Application**
  - **Graphs** are ideal for storing multiomics datasets due to their ability to model complex relationships.
  - However, traditional machine learning and deep learning architectures struggle with graph data due to its unstructured nature, where each node can have a variable number of neighbors without a fixed ordering.

Grid-like data



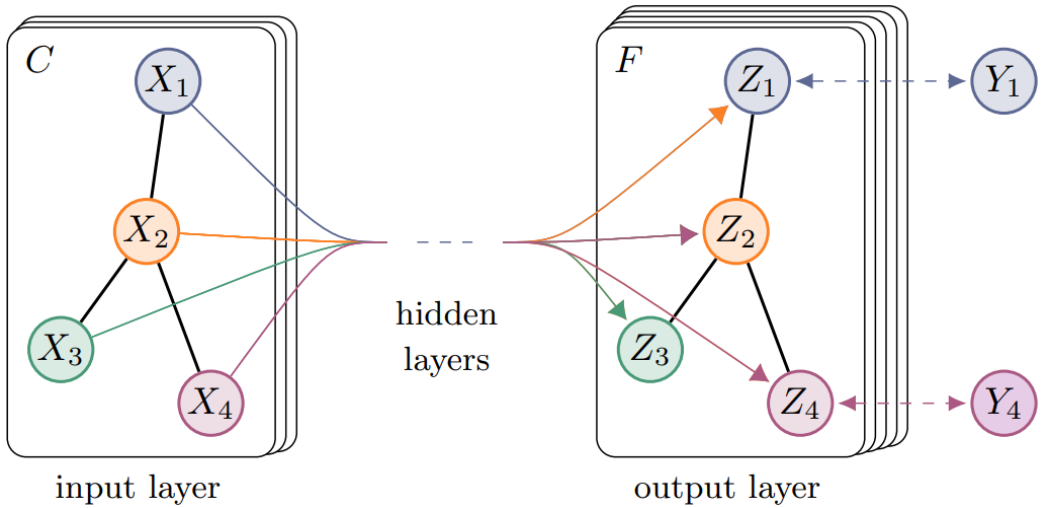
Graph data

- **Graph Neural Networks(GNN)** aggregate features from the local structure of graphs more effectively, making them suitable for large-scale and feature-rich datasets.

DOI:10.3390/s21144758

Introduction:

- **Graph Convolutional Networks (GCNs) in Cancer Research**
  - Among GNNs, **GCN** stand out for their modified aggregation technique that includes self-edges with normalization across neighbors, allowing for efficient feature aggregation.
  - GCNs have been increasingly applied to biological problems, including cancer subtype prediction and drug response prediction, showcasing their potential in leveraging graph-structured data for complex analyses.



https://doi.org/10.48550/arXiv.1609.02907

Introduction:

- **Single Network Application:**
  - Early applications of convolution methods to graph-structured data, including cancer subtyping, were limited to single networks. This approach often missed the integration of multiple data modalities, reducing the comprehensiveness of the analysis.

- **Case Studies Highlighting Limitations:**
  - An example is the application of GCNs for cancer type prediction using only gene expression datasets, which overlooks the richness of multiomics data.
  - **MOGONET**, a supervised multiomics integration framework using GCNs, operated on separate networks for mRNA expression, DNA methylation, and microRNA expression.

    → Did not integrate multiple network features, limiting its ability to fully leverage the available data for more accurate predictions.

Introduction:

- **Need for Comprehensive Integration:**
  - The limitations highlight the need for methods that can integrate features across multiple networks and data types.

  - **SUPREME** addresses these gaps by not only generating network-specific patient embeddings but also by integrating these embeddings with raw features from multiple omics data types.

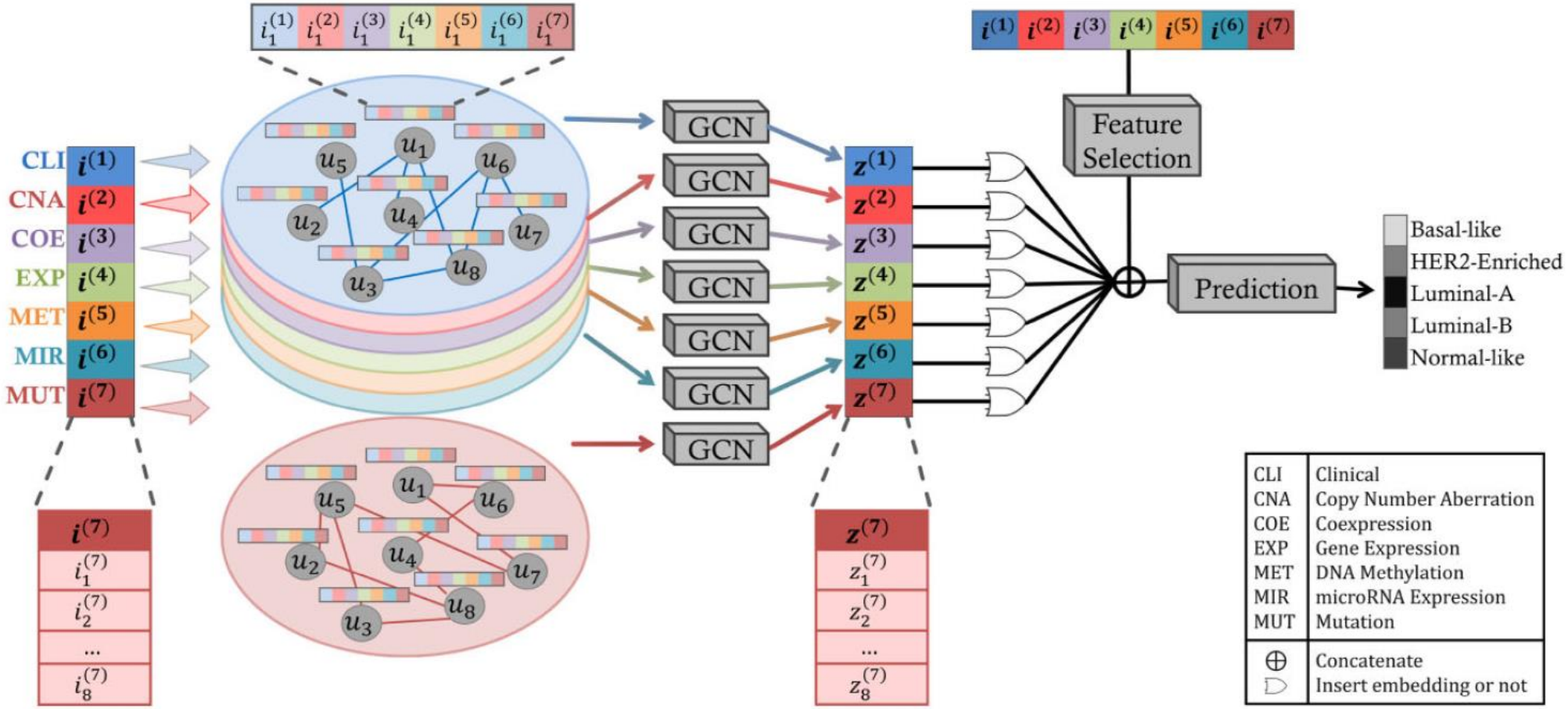
  - This comprehensive approach ensures that all available data is utilized, improving the accuracy of subtype predictions and capturing more nuanced signals within the data.

Introduction:

- **Performance:**
  - Demonstrated superior accuracy in cancer subtype prediction tasks compared to other integrative supervised methods.
  - Showed robust performance across multiple datasets, including TCGA and METABRIC.
  - SUPREME can differentiate the characteristics of cancer subtypes properly utilizing the multiple network relations and multiple data types
  - Revealed survival differences among cancer subtypes with greater significance than those identified by conventional gene expression-based methods.

Introduction:

- **SUPREME pipeline for breast cancer subtype prediction**



Materials and Methods

Materials and Methods:

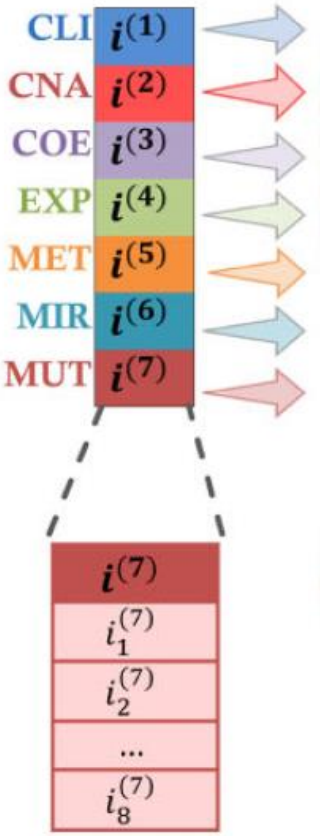
- **Process Steps:**

1. **Data Preparation:** Collection and preprocessing of data across various types for analysis.
2. **Feature Extraction:** Features are extracted from each data type to be utilized in further analysis.
3. **Similarity Network Generation:** Individual similarity networks are generated for each data type, incorporating features from all data types as node attributes.
4. **Network-Specific Node Embeddings:** GCN is applied to each network to generate network-specific node embeddings.
5. **Prediction:** Integrates individual network-specific embeddings and raw features for the final prediction task.

Materials and Methods: Data preparation

- **Datasets:**
  - **TCGA:** 1,022 breast tumor samples across seven data types (clinical, copy number aberration, coexpression, gene expression, DNA methylation, microRNA expression, mutation).
  - **METABRIC:** 1,699 breast tumor samples across five data types (clinical, copy number aberration, coexpression, gene expression, mutation).
  - **Combined TCGA and METABRIC:** 2,721 breast tumor samples across three data types (clinical, gene expression, mutation).

- **Ground Truth:** Utilizes PAM50 subtype labels (Basal-like, HER2-Enriched, Luminal-A, Luminal-B, Normal-like) for prediction tasks.



Materials and Methods: Data preparation

- **Additional Datasets:**
  - **IMDB Dataset (movie genre prediction):** Heterogeneous network with three node types (movie, actor, director) and two associations (movie-actor, movie-director), classifying movies into three genres: action, comedy, and drama.
  - **ACM Dataset (paper area prediction):** Heterogeneous network with three node types (paper, author, subject) and two associations (paper-author, paper-subject), classifying papers into three classes: database, wireless communication, and data mining.

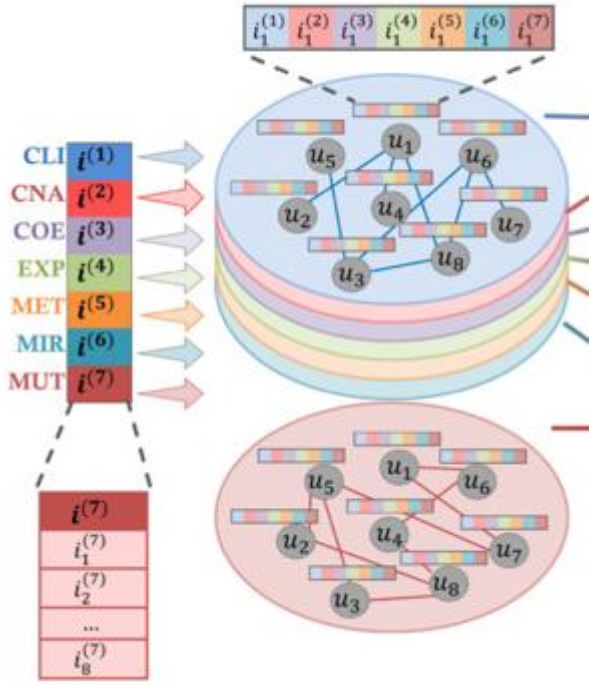
Materials and Methods: Data preparation

- **Table 1.** Number of features and samples for each dataset.
  BL: basal-like, HER2: HER2-Enriched, LA: luminal- A, LB: luminal-B, NL: normal-like

Dataset	Number of raw features	Number of samples
TCGA	3088	1022 samples: 172 BL (17%), 78 HER2 (8%), 538 LA (53%), 195 LB (19%), 39 NL (4%)
METABRIC	1761	1699 samples: 199 BL (12%), 220 HER2 (13%), 679 LA (40%), 461 LB (27%), 140 NL (8%)
Combined (TCGA+ METABRIC)	1229	2721 samples: 371 BL (14%), 298 HER2 (11%), 1217 LA (45%), 656 LB (24%), 179 NL (7%)
IMDB	3066	4278 samples: 1135 (27%), 1584 (37%), 1559 (36%)
ACM	1870	3025 samples: 1061 (35%), 965 (32%), 999 (33%)

Materials and Methods: Feature extraction & network generation

- **Breast cancer subtyping**
  - Utilized **Boruta algorithm** for feature selection across 7 data types from **TCGA**, 5 from **METABRIC**, and 3 from **combined** datasets.
  - Selected features are used to compute patient similarities, serve as node attributes in **similarity networks**, and integrate as raw features for **prediction**.
  - Patient similarities calculated using **Pearson correlation**, **Gower metric**, and **Jaccard distance** for various data types including gene expression and clinical features.
  - Constructs unweighted patient similarity networks using a defined number of top edges: 2,500 for TCGA, 4,500 for METABRIC, and 7,000 for combined data.



Materials and Methods: Feature extraction & network generation

- Used **Pearson correlation** for datatypes with continuous values

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Used **Jaccard Similarity** for binary datatype (mutation datatype)

$$Jaccard(U, V) = \frac{|U \cap V|}{|U \cup V|}$$

- Used **Gower metric** for mixed datatype (clinical datatype), containing continuous, binary, and categorical at the same time.
  → Numerical variables: Manhattan distance
  → Categorical variables: Dice distance

Materials and Methods: Feature extraction & network generation

- **IMDB Dataset (Movie Genre Prediction):**
  - No feature selection was applied; node features were directly used from previous processing.
  - Generated **2 movie similarity networks** using meta-paths: movie-director-movie (17,446 edges) and movie-actor-movie (85,358 edges), indicating similarity based on shared directors or actors.

- **ACM Dataset (Paper Area Prediction):**
  - Similarly, no feature selection was applied, with node features used as processed in prior studies.
  - Created **2 paper similarity networks** via meta-paths: paper-author-paper (29,281 edges) and paper-subject-paper (2,210,761 edges), with similarity defined by shared authors or subjects.

Materials and Methods: Node embedding generation

- **Embedding Generation:** SUPREME creates network-specific node embeddings that reflect both the network topology and the individual node features, ready for downstream machine learning tasks.

- **GCN Model Utilization:** Employs the GCN model by *Kipf and Welling*, which incorporates self-edges in convolution and normalizes the sum of aggregated features from neighbors.
  - This approach enables learning from the data by considering the one-hop local neighborhood and encoding the network's local topology.

Materials and Methods:

- **Layered Network Learning:**
  - Stacked layers allow for the recursive diffusion of neighborhood features, extending beyond the immediate one-hop neighborhood to capture broader network contexts.

- **Graph Representation:**
  - Graphs are defined as **undirected**, with $G = (V, E)$ where V represents nodes and E represents edges. Nodes are interconnected based on associations within the graph, ensuring $(v_i, v_j) \in E$ implies a bidirectional association $(v_j, v_i) \in E$

Materials and Methods:

- **GCN Input and Iteration:**
  - Inputs include a feature matrix X of dimensions n x k (where n is the number of nodes, and k is the feature size) and an adjacency matrix A with self-edges defined as:

$$\mathcal{A}[i, j] = \begin{cases} 1 & \text{if } (v_i, v_j) \in \mathcal{E} \text{ or } i = j \\ 0 & \text{otherwise} \end{cases}$$

  - Iterative updates follow the formula:

$$\mathcal{H}^{(l+1)} = \sigma\left(\mathcal{D}^{-\frac{1}{2}}\mathcal{A}\mathcal{D}^{-\frac{1}{2}}\mathcal{H}^{(l)}\mathcal{W}^{(l)}\right)$$

    - where D is a degree matrix,
    - $H^{(l)}$ represents the activation matrix for layer l,
    - $W^{(l)}$ is the layer's weight matrix, and
    - σ denotes the activation function.

Materials and Methods:

- **Network Configuration:**
  - Seven patient similarity networks created, each derived from a distinct data type.
  - Nodes represent breast cancer patients, connected based on similarity metrics from their respective data types.

- **Example Network:**
  - Gene expression-derived patient similarity network (G) connects patients with similar gene expression profiles.
  - Node features in G include combined features from all seven data types.

Materials and Methods:

- **Feature Matrix and Neighborhood:**
  - Features for a node vi are represented as $\mathrm{x_i} \in R^k$, where k is the total number of features.
  - The feature matrix $\mathrm{X} \in R^{n \times k}$ combines features from all patients.
  - The one-hop local neighborhood of a node, Ni, includes nodes directly connected to vi.

- **Feature Aggregation and GCN Layers:**
  - Feature aggregation employs a scaled adjacency matrix

$$\mathcal{A}' = \mathcal{D}^{-\frac{1}{2}} \mathcal{A} \mathcal{D}^{-\frac{1}{2}}.$$

  - A 2-layer GCN model outputs

$$\mathcal{Z} = \operatorname{softmax}\left(\mathcal{A}' \operatorname{ReLU}\left(\mathcal{A}' \mathcal{X} \mathcal{W}^{(1)}\right) \mathcal{W}^{(2)}\right)$$

Materials and Methods:

- **Prediction and Optimization:**
  - Aims to predict five classes corresponding to breast cancer subtypes: Basal-like, Luminal-A, Luminal-B, HER2-Enriched, and Normal-like.
  - Utilizes **cross-entropy error** for loss calculation and **Adam optimization** for gradient descent.
  - Incorporates dropout in the first GCN layer to prevent overfitting and applies early stopping with a patience of 30 epochs to ensure a minimum of 200 epochs for model training.

Materials and Methods:

- **Stratified Splitting:** Samples divided into training, validation, and test sets, maintaining subtype label ratios across splits.
  - **Test set** reserved solely for final evaluation.
  - **Training** and **validation sets** randomly selected in a stratified manner for each run.

- **Breast Cancer Subtyping Splits:**
  - 20% of samples designated as the test set.
  - Of the remaining 80%, 60% used for training and 20% for validation.

- **IMDB and ACM Datasets:**
  - Followed data splits specified in previous studies.

Materials and Methods:

- **Hyperparameter Tuning:**
  - Utilized **macro-averaged F1 score** (macro F1) for evaluation.
  - Repeated evaluation 10 times for each combination of hyperparameters (hidden layer size, learning rate).
  - Best hyperparameter combination chosen based on the median macro F1 score on validation data.

- **Model Generation:**
  - Produced 7 different GCN models for TCGA data, 5 for METABRIC data, 3 for combined data, and two each for ACM and IMDB data based on the methodology.
  - Final models used to extract network-specific patient embeddings for downstream prediction tasks.

Materials and Methods:

- **Combining Embeddings:** Node embeddings from various data types were concatenated with raw features for model training, resulting in multiple models based on data type combinations.

- **Models by Dataset:** Generated 127 SUPREME models for TCGA, 31 for METABRIC, 7 for combined TCGA+METABRIC, and 3 each for ACM and IMDB datasets.

- **ML Methods and Selection:** Tested with XGBoost, SVM, RF, and MLP; MLP chosen for superior performance across all datasets.
- **Hyperparameter Tuning:** Similar process to GCN, focusing on prediction model parameters with the best set determined by median macro F1 score on validation data.
- **Final Evaluation:** Final models evaluated on test data over 10 runs, using metrics like macro F1, weighted-average F1 score, and accuracy based on these runs' median.

Results

Results:

- **Comparison with Other Tools:** Evaluated against seven other cancer prediction tools and baseline methods: DeepCC, GCNC, MOGONET, MLP, RF, SVM, and XGBoost.

- **Baseline Methods Integration:** ML-based methods (MLP, RF, SVM, XGBoost) utilized only raw features from selected data combinations for prediction.

- **MOGONET and GCNC:**
  - MOGONET employed GCN for multiomics data, using datatype-specific embedding predictions.
  - GCNC applied GCN on gene expression data through either a protein-protein interaction (PPI) network or a coexpression network.
  - DeepCC relied solely on gene expression data, transforming pathway activity via an MLP model.

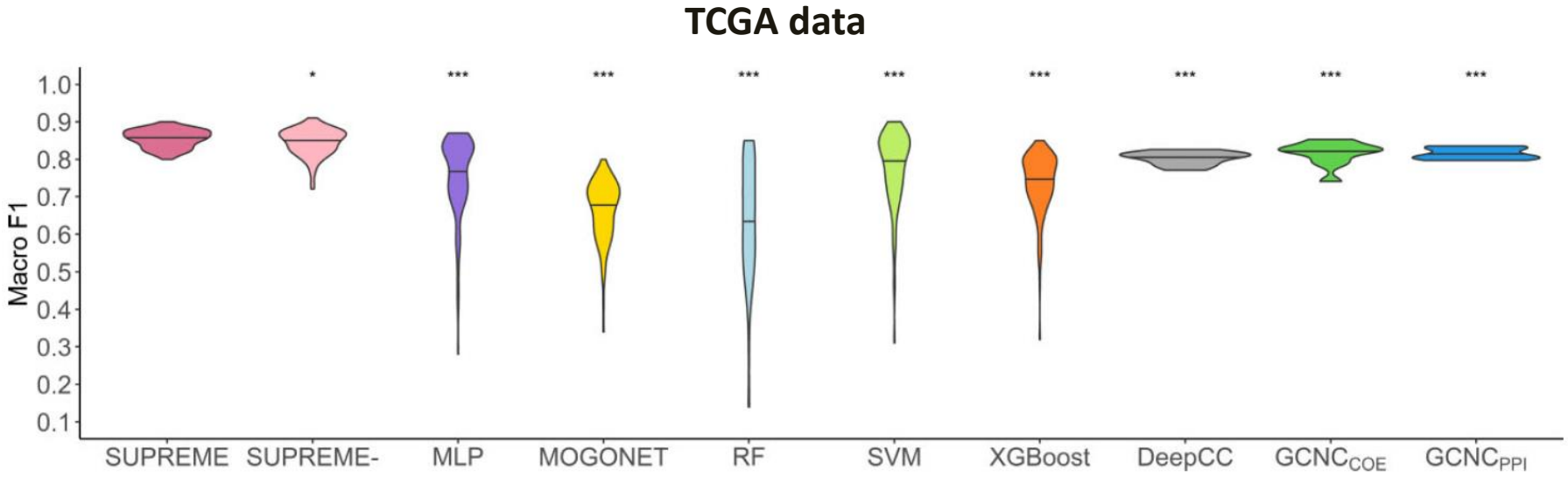
Results:

- **SUPREME Variants:** Tested SUPREME both with and without (SUPREME-) integrating raw features with patient embeddings to assess the impact of raw feature integration.

- **Model Testing Across Data Combinations:** Ran SUPREME, SUPREME-, and other models across all available data type combinations. Notably, MOGONET was limited to fewer than five data types due to extended processing times, resulting in only 31 model variations for TCGA, whereas all models were tested for METABRIC and combined datasets.

Results: SUPREME outperformed the cancer subtype prediction tools and baseline methods

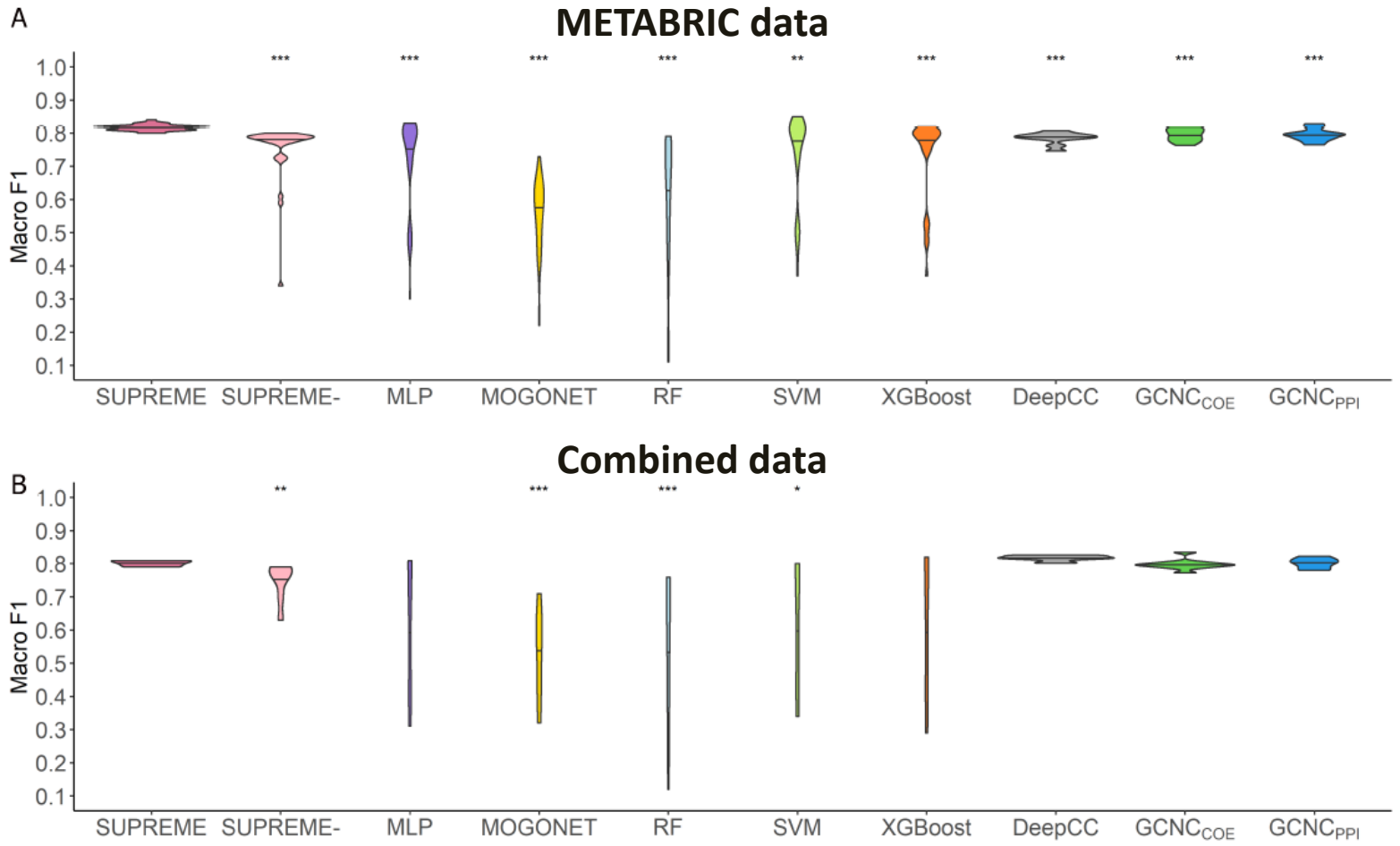
- **Violin plot** of macro F1 scores obtained from all different combinations of datatypes as compared to the cancer subtype prediction tools and baseline supervised methods



- SUPREME and its variant without raw feature integration (SUPREME-) surpassed all other multiomics integration methods across three datasets in macro F1, accuracy, and weighted F1 scores.

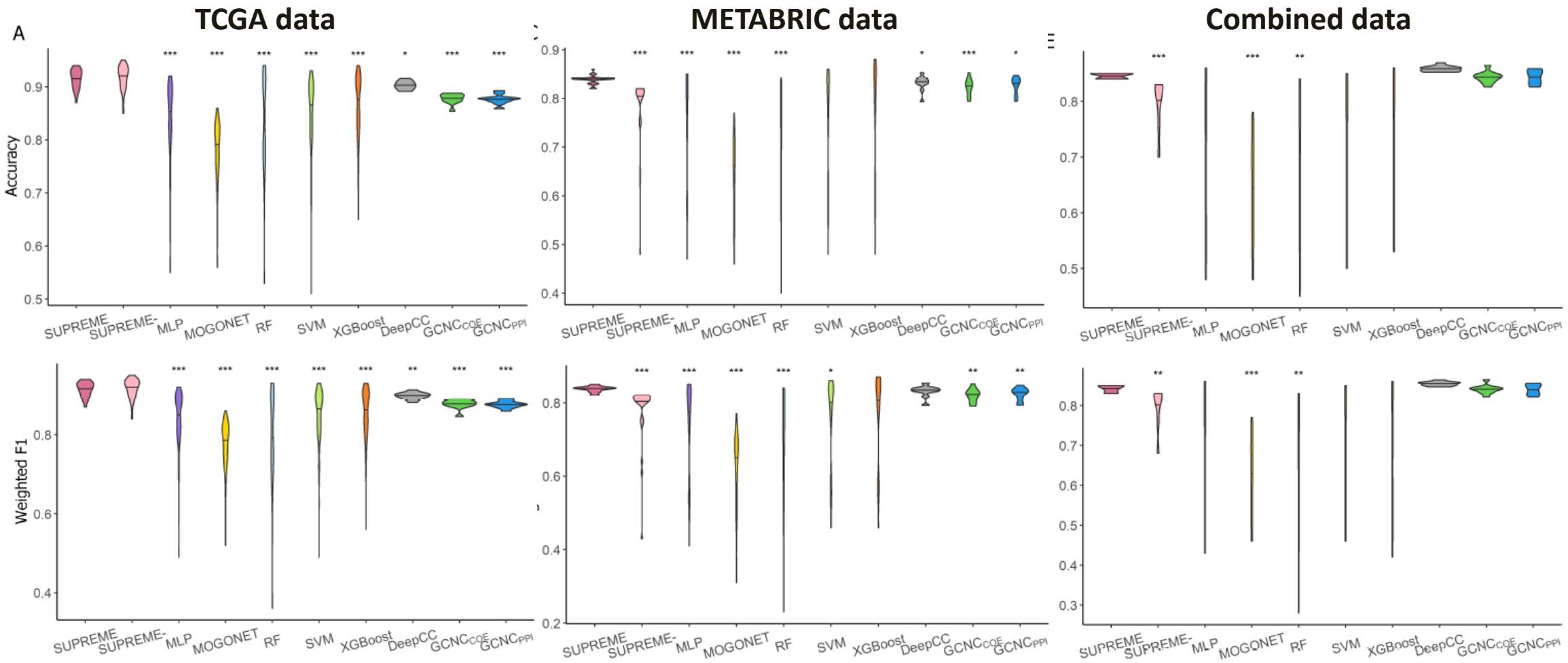
Results: SUPREME outperformed the cancer subtype prediction tools and baseline methods

- **Violin plot** of macro F1 scores obtained from all different combinations of datatypes as compared to the cancer subtype prediction tools and baseline supervised methods



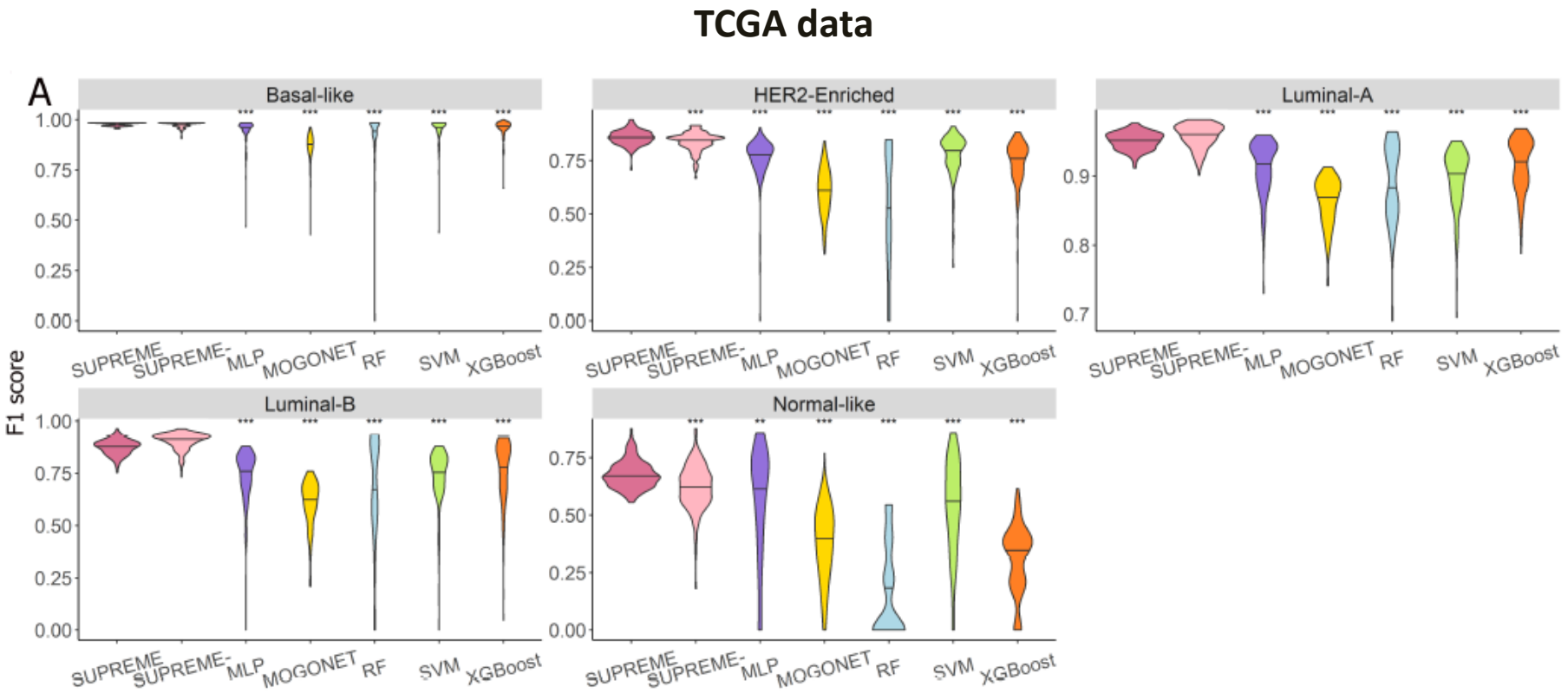
Results: SUPREME outperformed the cancer subtype prediction tools and baseline methods

- **Violin plot** of accuracies and weighted F1 scores



Results: SUPREME outperformed the cancer subtype prediction tools and baseline methods

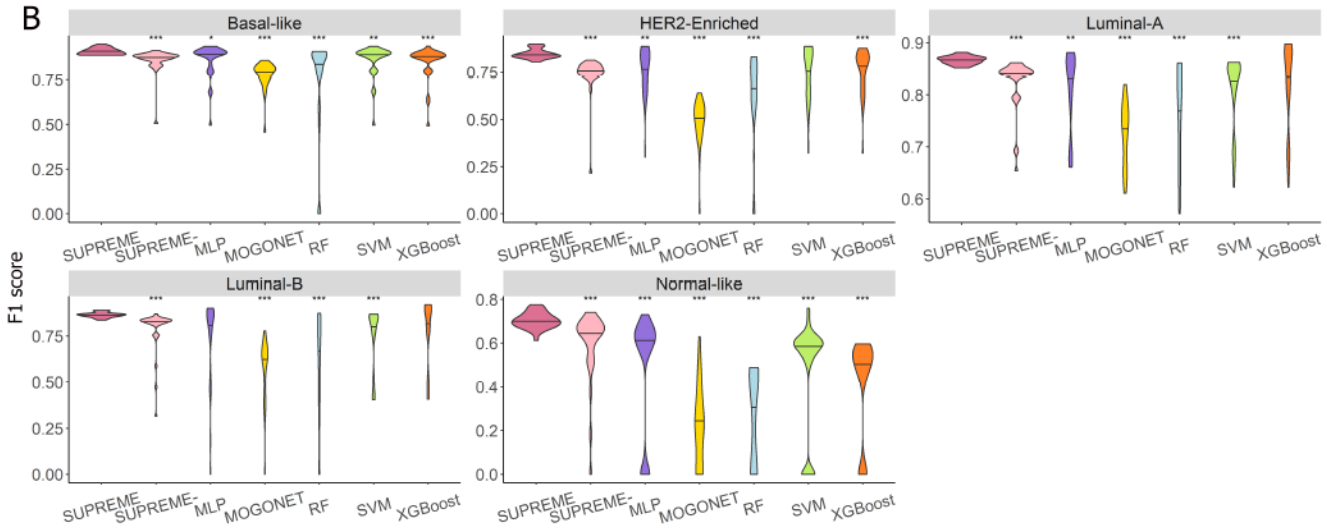
- **Subtype-specific F1 scores.**
  - Checked the subtype-specific F1 scores, and had consistent and higher performance across all sub- types, mostly having significant differences



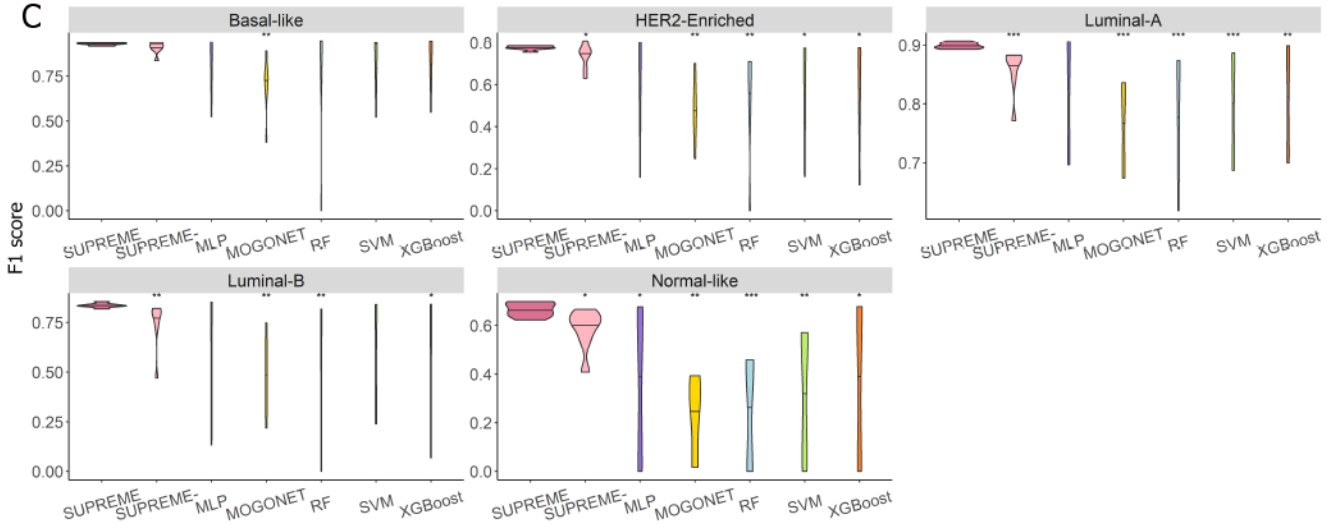
Results: SUPREME outperformed the cancer subtype prediction tools and baseline methods

- **Subtype-specific F1 scores.**

METABRIC data



Combined data



Results: SUPREME had consistently high performance even with single models

- **Single Model Analysis:** Investigated SUPREME's performance using models generated from only one data type, referred to as "single model."

Single model results on TCGA data

Macro F1 scores

Method	CLI	CNA	COE	EXP	MET	MIR	MUT
SUPREME	0.68 ± 0.04	**0.80 ± 0.03**	0.76 ± 0.04	**0.84 ± 0.02**	**0.79 ± 0.03**	0.73 ± 0.02	**0.75 ± 0.03**
SUPREME-	**0.72 ± 0.02**	0.77 ± 0.02	**0.77 ± 0.04**	0.77 ± 0.05	**0.79 ± 0.04**	0.70 ± 0.02	0.74 ± 0.04
MLP	0.46 ± 0.07	0.53 ± 0.04	0.59 ± 0.02	0.82 ± 0.03	0.69 ± 0.04	**0.74 ± 0.04**	0.28 ± 0.06
MOGONET	0.41 ± 0.01	0.52 ± 0.01	0.57 ± 0.01	0.75 ± 0.01	0.61 ± 0.03	0.71 ± 0.03	0.34 ± 0.01

Weighted F1 scores

Method	CLI	CNA	COE	EXP	MET	MIR	MUT
SUPREME	0.80±0.02	**0.87±0.02**	0.85±0.01	**0.89±0.01**	**0.87±0.02**	**0.83±0.01**	**0.83±0.01**
SUPREME-	**0.83±0.01**	0.86±0.01	**0.86±0.01**	0.86±0.02	**0.87±0.02**	0.82±0.01	**0.83±0.01**
MLP	0.62±0.04	0.68±0.03	0.77±0.01	0.88±0.01	0.79±0.02	**0.83±0.02**	0.49±0.06
MOGONET	0.58±0.01	0.68±0.01	0.75±0.01	0.82±0.00	0.73±0.01	0.78±0.02	0.53±0.00

Accuracies

Method	CLI	CNA	COE	EXP	MET	MIR	MUT
SUPREME	0.81±0.01	**0.87±0.02**	**0.86±0.01**	**0.89±0.01**	**0.88±0.02**	0.83±0.01	**0.83±0.01**
SUPREME-	**0.83±0.02**	**0.87±0.01**	**0.86±0.01**	0.87±0.01	0.87±0.02	0.82±0.01	**0.83±0.01**
MLP	0.71±0.03	0.70±0.03	0.79±0.01	0.88±0.01	0.80±0.02	**0.84±0.02**	0.55±0.03
MOGONET	0.62±0.02	0.70±0.01	0.77±0.01	0.82±0.00	0.74±0.00	0.78±0.02	0.56±0.01

Results: SUPREME had consistently high performance even with single models

- **Single model results on METABRIC and combined data**

METABRIC data

Macro F1 scores

Method	CLI	CNA	COE	EXP	MUT
SUPREME	**0.49±0.01**	0.45±0.02	**0.75±0.01**	0.78±0.02	**0.36±0.03**
SUPREME-	**0.49±0.03**	0.44±0.01	**0.75±0.01**	**0.82±0.01**	0.33±0.04
MLP	0.46±0.02	**0.46±0.02**	0.71±0.02	0.81±0.02	0.30±0.03
MOGONET	0.41±0.02	0.42±0.01	0.65±0.03	0.73±0.01	0.22±0.00

Weighted F1 scores

Method	CLI	CNA	COE	EXP	MUT
SUPREME	**0.54±0.01**	0.52±0.02	**0.78±0.01**	0.81±0.02	**0.45±0.03**
SUPREME-	**0.54±0.04**	0.51±0.01	**0.78±0.01**	**0.84±0.01**	0.41±0.04
MLP	0.50±0.03	**0.53±0.01**	0.76±0.02	**0.84±0.01**	0.41±0.04
MOGONET	0.47±0.01	0.50±0.01	0.69±0.01	0.77±0.01	0.31±0.00

Accuracies

Method	CLI	CNA	COE	EXP	MUT
SUPREME	**0.58±0.01**	**0.56±0.02**	**0.79±0.01**	0.82±0.02	**0.49±0.01**
SUPREME-	**0.58±0.01**	0.54±0.01	0.78±0.01	0.84±0.01	**0.49±0.01**
MLP	**0.58±0.01**	**0.56±0.01**	0.75±0.02	**0.85±0.01**	0.47±0.01
MOGONET	0.52±0.02	0.52±0.01	0.70±0.01	0.77±0.01	0.46±0.00

Combined data

Macro F1 scores

Method	CLI	EXP	MUT
SUPREME	0.43±0.01	**0.81±0.01**	**0.33±0.03**
SUPREME-	0.43±0.01	0.80±0.01	**0.33±0.01**
MLP	**0.44±0.01**	0.80±0.01	0.31±0.03
MOGONET	0.41±0.00	0.71±0.01	0.32±0.01

Weighted F1 scores

Method	CLI	EXP	MUT
SUPREME	0.53±0.01	**0.85±0.01**	**0.47±0.03**
SUPREME-	0.52±0.02	0.84±0.01	0.46±0.01
MLP	**0.54±0.01**	0.84±0.01	0.43±0.02
MOGONET	0.50±0.00	0.77±0.01	0.46±0.01

Accuracies

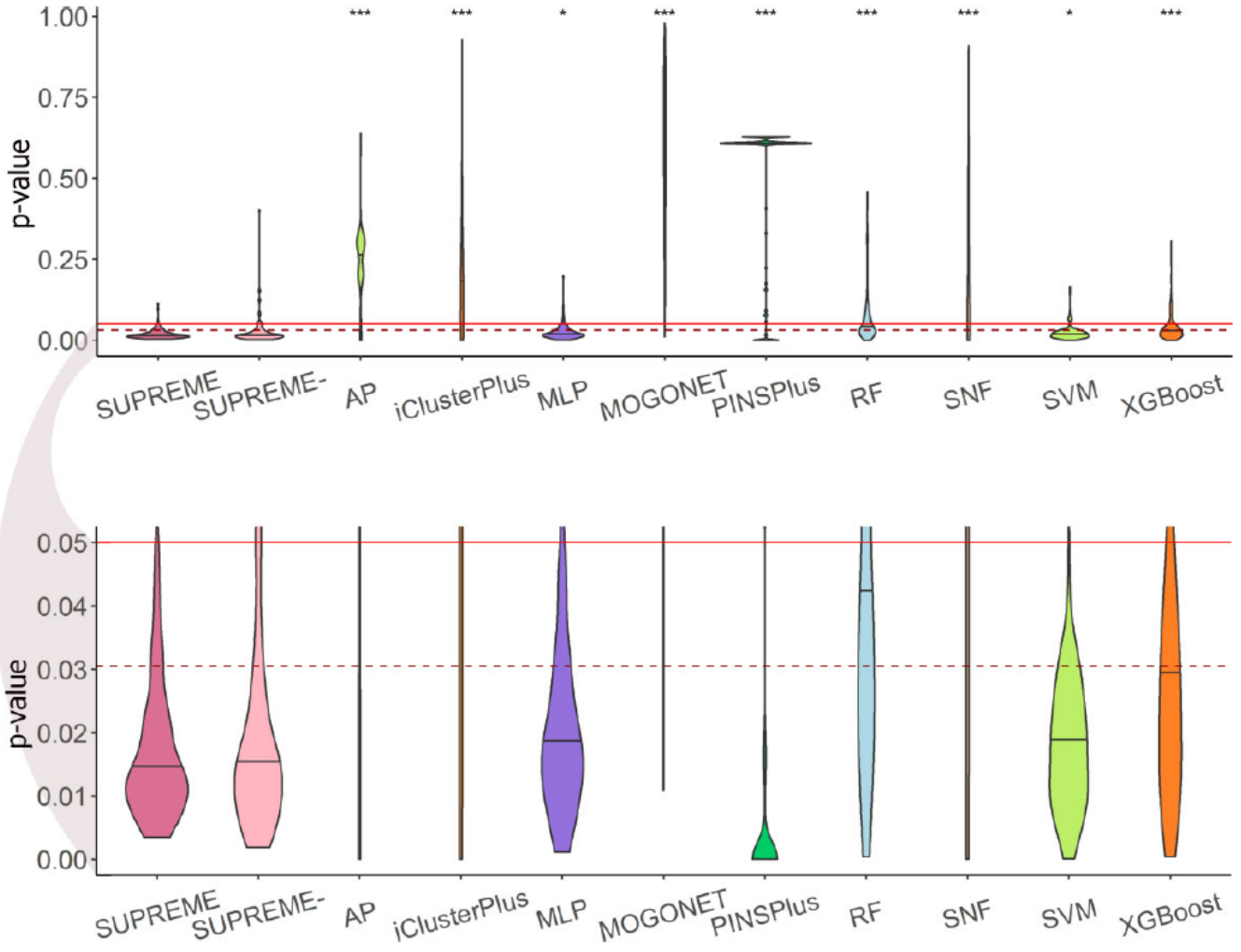
Method	CLI	EXP	MUT
SUPREME	**0.58±0.00**	**0.85±0.01**	**0.51±0.02**
SUPREME-	**0.58±0.01**	0.84±0.01	**0.51±0.01**
MLP	**0.58±0.01**	**0.85±0.01**	0.48±0.02
MOGONET	0.55±0.00	0.78±0.01	0.48±0.01

Results: SUPREME's Impact on Survival Differences in Cancer Subtyping

- **Survival Analysis Strategy:**
  - Only applied to the results on **TCGA data** where patient survival data were available.
  - Utilized predicted subtype labels across different data modality combinations for survival analysis.
  - Comparing SUPREME to both supervised and state-of-the-art unsupervised cancer subtyping tools, including **iClusterPlus**, **SNF**, **PINSPlus**, and **affinity propagation (AP)** clustering.

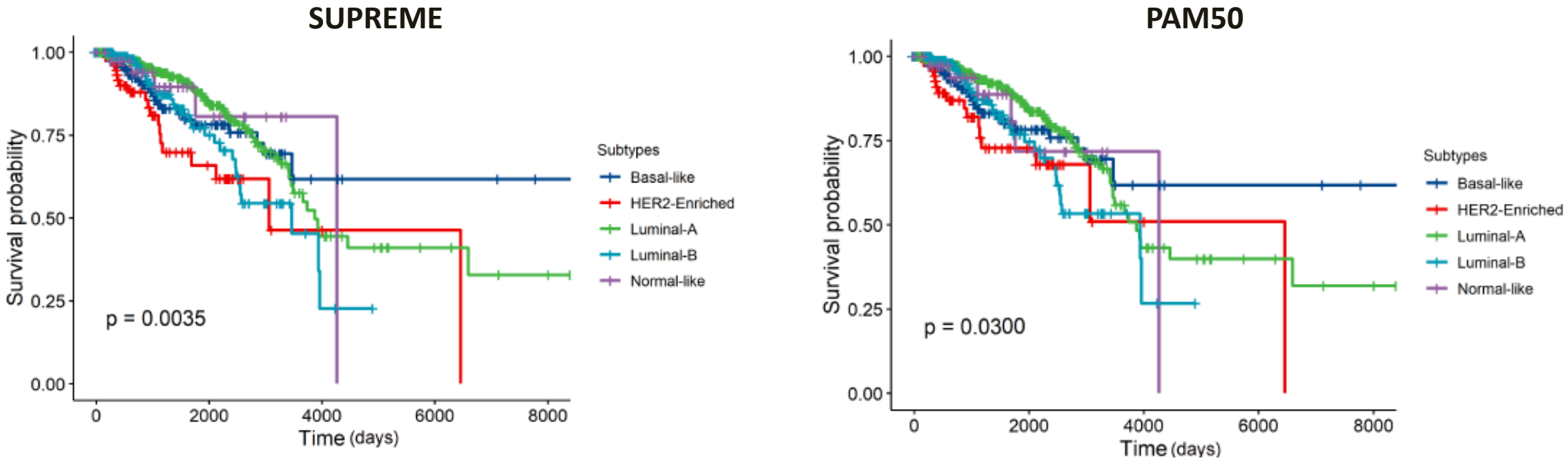
Results: SUPREME's Impact on Survival Differences in Cancer Subtyping

- **Survival analysis results violin plot**



Results: SUPREME's Impact on Survival Differences in Cancer Subtyping

- **Kaplan-Meier plots** of two cases: SUPREME & PAM50 labels

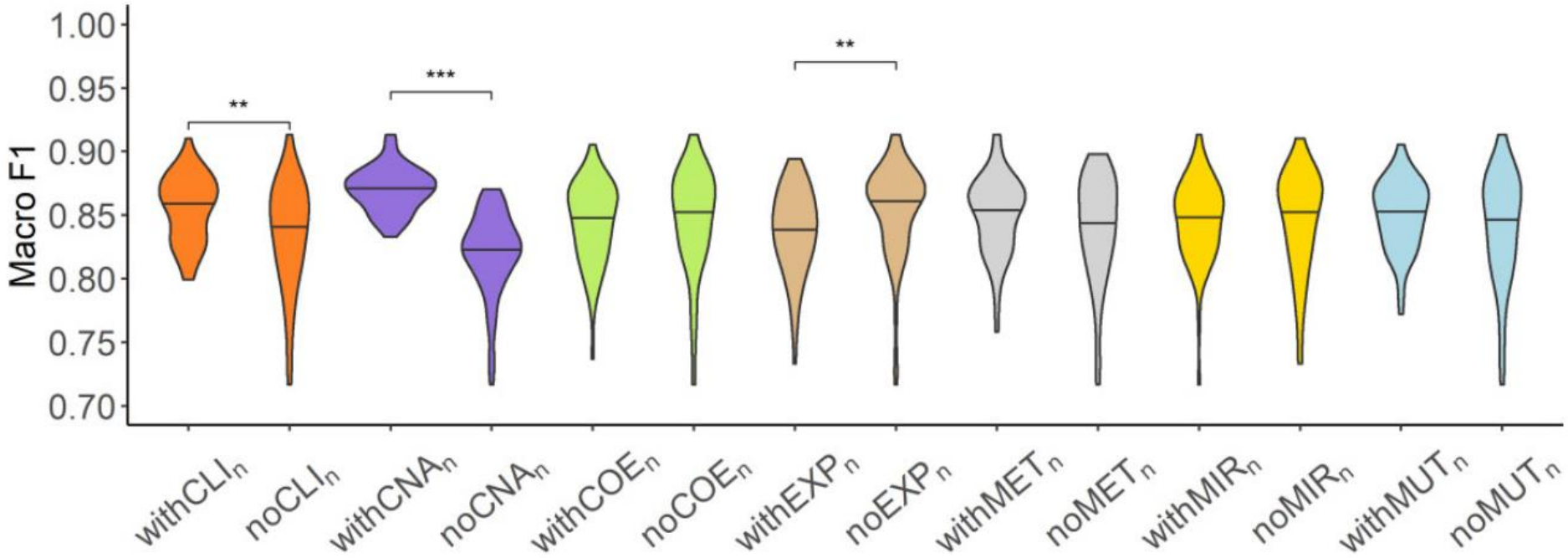


- SUPREME had a more significant survival difference than the survival difference between ground truth (i.e., PAM50) labels

Results: Feature / omics importance analysis

- **Impact of network-specific patient embeddings.**
  - Assessed each data type's impact on model performance by comparing models with and without a specific patient embedding from data type X, using SUPREME- (excluding raw feature integration) to focus solely on patient embeddings' effects.

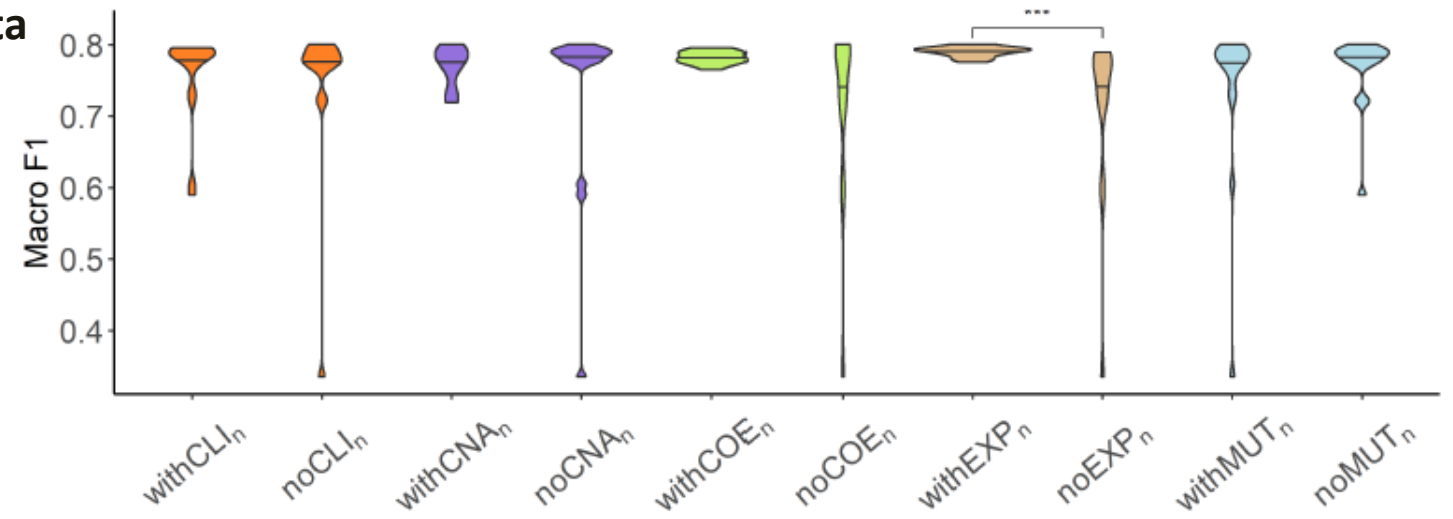
TCGA data



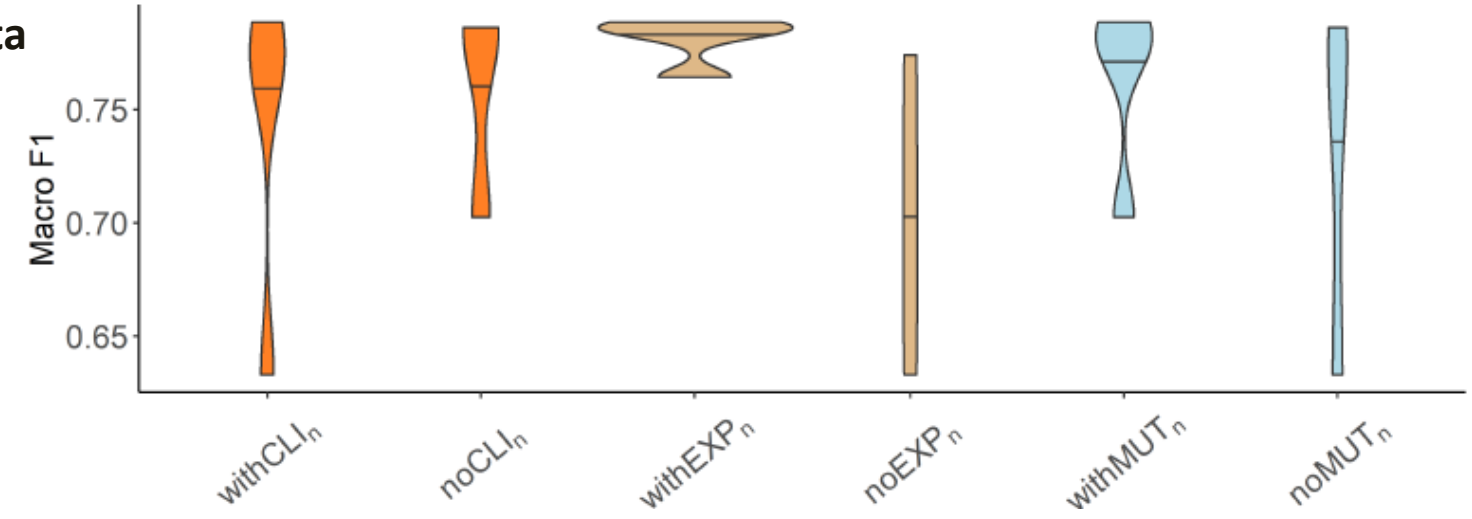
Results: Feature / omics importance analysis

- **Impact of network-specific patient embeddings.**

METABRIC data



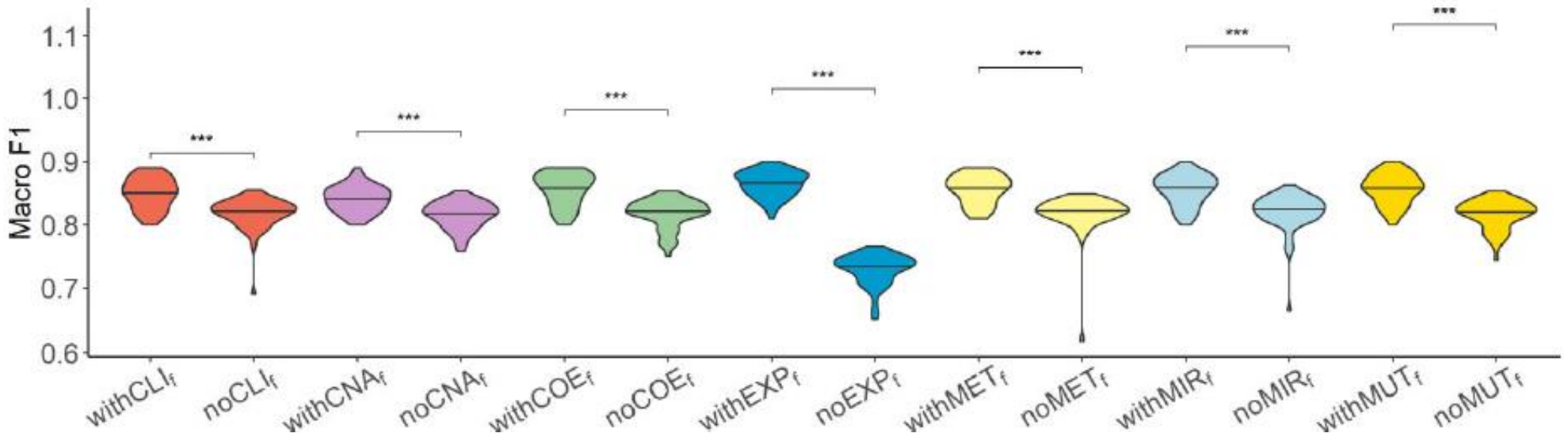
Combined data



Results: Feature / omics importance analysis

- **Impact of features from each datatype.**
  - SUPREME's performance was evaluated by excluding features from each datatype (Y) separately, creating models without Y-specific features ($no\ Y_f$) for patient similarity networks and subtype prediction.

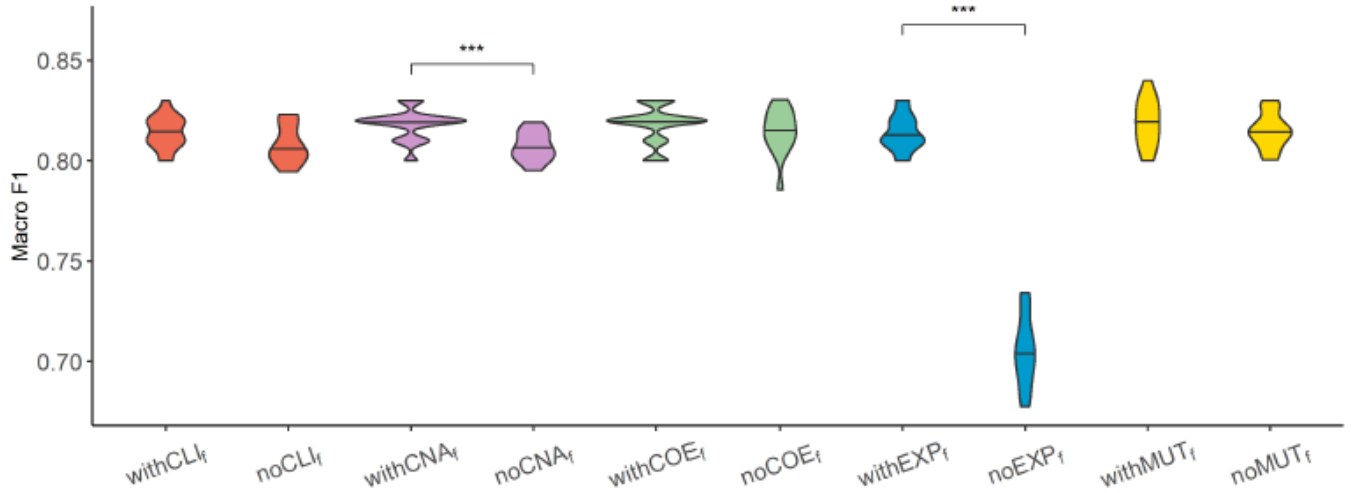
TCGA data



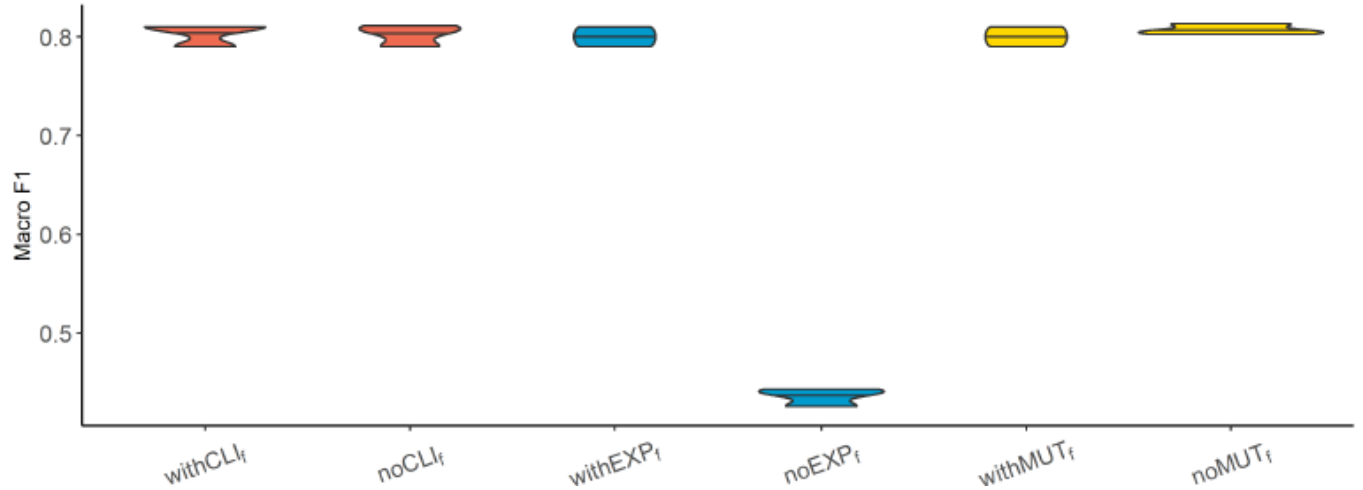
Results: Feature / omics importance analysis

- **Impact of features from each datatype.**

METABRIC data



Combined data



Results: Ablation studies

- Compared **SUPREME** with its variations when some steps were skipped to assess their importance

Comparison/Section	Measures impact of
SUPREME vs. SUPREME-	Raw feature integration
SUPREME vs. MLP	GCN utilization
Single model section	The used method with only one datatype
SUPREME vs. MLP in single model section	GCN utilization with only one datatype
This section	Node features

Discussion

Discussion:

- **SUPREME** is an integrative approach using GCNs on multiple similarity networks with multi-modal node features, outperforming other methods on various prediction tasks and cancer subtype differentiations.

- It demonstrated robustness, consistency, and significant survival differences between predicted subtypes, offering ensemble subtype labels with high support.

- Ablation study showed the importance of gene expression features, patient embeddings, and datatype-specific embeddings in improving model performance.

Discussion:

- **Limitations:**
  - Variations in predictive performance across different datasets, particularly noted in mutation-based models
  - These discrepancies were mainly due to the sparse nature of the binary mutation features.

- **Potential Extensions:**
  - Incorporating attention mechanisms to weigh contributions from different data types and neighborhoods more dynamically, enhancing the model's ability to focus on the most relevant information for subtype prediction.
  - Expanding SUPREME to include regulatory relations, like competing endogenous RNA (ceRNA) interactions, to enrich patient similarity networks with gene regulatory interactions and complex patient relations, offering deeper insights into cancer biology.

Thank you