Comp683: Computational Biology

Lecture 17

March 29, 2025

Today

- Graph Neural Networks vs Label Propagation vs LP + Correct and Smooth
- Examples of circumstances where spatial context is and is not helpful revealed through work with GNNs

Review Question

- What kind of model did LEPAH use?
- After defining the Ω in LEAPH, what was the optimization problem formulated to accommodate spatial information?

Ω Clean-Up

$$\min_{\Omega} -\sum_{i=1}^{N}\sum_{j=1}^{M}\Omega_{ij}\log_2\left(\Omega_{ij}\right) + \lambda\sum_{(m,n)} w_{mn}||\Omega_m - \Omega_n||_2$$

- w_{jk} is a weight, calculate as the reciprocal of distance between cells j and k in the image
- The first term is basically an entropy term of ownership confidence
- The second term is promoting spatial coherence.
- λ controls the tradeoff between spatial coherence and membership confidence.

Tradeoffs

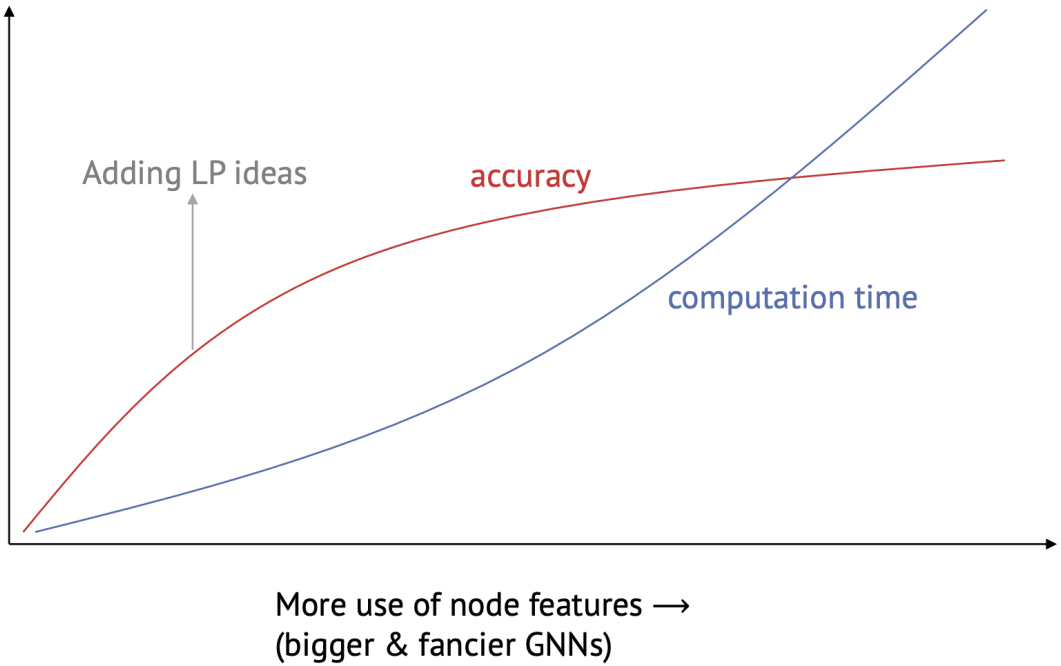


Figure: from `https://www.cs.cornell.edu/~arb/slides/2021-03-12-northeastern.pdf`

Correct and Smooth Approach

- The goal is to compare how a couple of simple methods/intuition can be strung together can be used to classify nodes
- The main idea is to start with a cheap base prediction based on node features (e.g. attributes or coordinates of a spectral embedding), and clean up graph structure through label propagation (**correct and smooth**).

Three Step Process

1. A base prediction made with node features that ignores the graph structure (e.g. with a linear model)
2. A correction step which propagates uncertainties from the training data across the graph to correct the base prediction
3. A smoothing of the predictions over the graph.

Overview of Correct and Smooth Approach

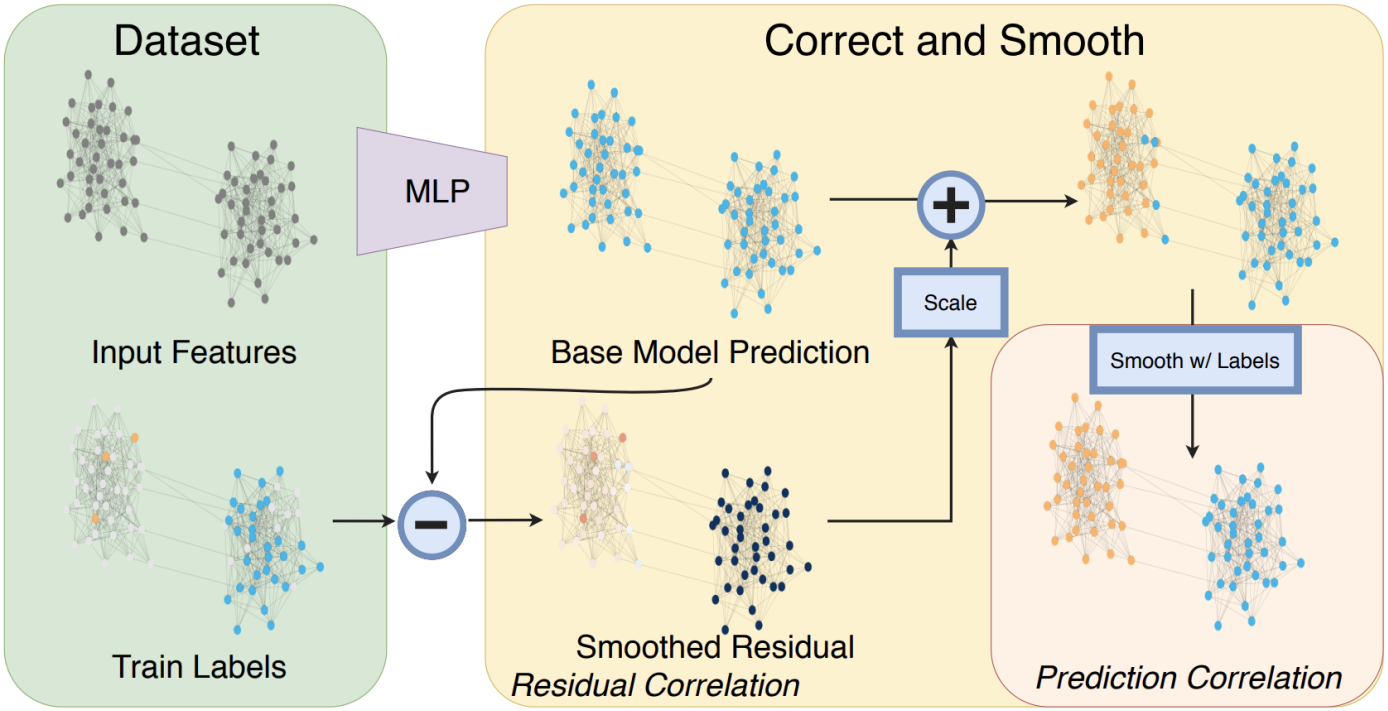


Figure: from Huang *et al.* ICLR. 2021

Notation Preliminaries

- Let there be n nodes.
- Assume we have a feature vector for each node, such that node features are encoded in an $n \times p$ matrix, X.
- Similarly, let A be the adjacency matrix of the graph
- Split nodes into labeled (L) and unlabeled (U) sets
- Define an $n \times c$ matrix, Y with a binary indicator for whether node i is in class c.

Simple Base Predictor

Given the matrix of **features** for each node, X and labels, Y, train a simple model to minimize,

$$\sum_{i \in L_t} \ell\left(f\left(x_i\right), y_i\right)$$

- ℓ is some loss
- Here L_t denotes the set of labeled training nodes
- Specify a matrix, Z containing these base predictions.

Error Correlation - Label Spreading Technique

- The intuition is that errors are expected to be correlated across edges in the graph. Hence, spread uncertainty across the edges.

Define and error matrix, $E \in \mathbb{R}^{n \times c}$ as,

$$E_{L_t,:} = Y_{L_t,:} - Z_{L_t,:}, \quad E_{L_v,:} = 0, \quad E_{U,:} = 0$$

This means that the only non-zero entries are those that correspond to labeled training nodes! These entries represent **residuals**.

Smooth the Error Using a Label Spreading Technique

The errors are smoothed as follow with a label spreading technique,

$$\hat{E} = \underset{W \in \mathbb{R}^{n \times c}}{\arg\min}\ \text{trace}\left(W^T(I - S)W\right) + \mu\|W - E\|_F^2$$

- S is the normalized adjacency matrix, $D^{-1/2}AD^{-1/2}$
- The first term encourages smoothness of the error over the graph
- The second term keeps W close to the initial estimate of error, E.

Our Friend Smoothness and Quadratic Form

We keep seeing the quadratic form come up if we are talking about smoothness. Reminder that,

$$\text{trace}(W^T(I - S)W) = \sum_j w_j^T(I - S)w_j$$

.

- $W \in \mathbb{R}^{n \times c}$

Solution

Given

$$\hat{E} = \underset{W \in \mathbb{R}^{n \times c}}{\arg\min}\ \text{trace}\left(W^T(I - S)W\right) + \mu\|W - E\|_F^2$$

it was previously shown that the solution can be obtained through the following iteration,

$$E^{(t+1)} = (1 - \alpha)E + \alpha S E^{(t)}$$

The quickly converges to $\hat{E}$ and therefore gives corrected predictions as,

$$Z^r = Z + \hat{E}$$

Smoothing Final Predictions with Prediction Correlation

- The next assumption to be used for correction is that adjacent nodes in the graph are likely to have similar labels (e.g. homophily)
- Another round of label propagation will be used to encourage smoothness over distribution of labels.

Starting with the best guess of the labels, H, with $H_{L_t,:} = Y_{L_t,:}$ and $H_{L_v \cup U,:} = Z^{(r)}_{L_v \cup U,:}$, propagate labels as,

$$H^{(t+1)} = (1 - \alpha)H + \alpha S H^{(t)}$$

Final Prediction

The following has now been applied

- Base prediction
- Residual correction
- Label smoothing

After convergence of $H^{(t+1)} = (1 - \alpha)H + \alpha SH^{(t)}$, get a final prediction, $\hat{Y} \in \mathbb{R}^{n \times c}$, and assign node to the class with the max predicted probability.

Results

Datasets	Classes	Nodes	Edges	Parameter Δ	Accuracy Δ	Time (s)
Arxiv	40	169,343	1,166,243	−84.90%	+0.26	12 (+90)
Products	47	2,449,029	61,859,140	−93.47%	+1.74	171 (+2959)
Cora	7	2,708	5,429	−98.37%	+1.09	< 1 (+7)
Citeseer	6	3,327	4,732	−89.68%	−0.69	< 1 (+7)
Pubmed	3	19,717	44,338	−96.00%	−0.30	< 1 (+14)
Email	42	1,005	25,571	−97.89%	+4.33	43 (+17)
Rice31	10	4,087	184,828	−99.02%	+1.39	39 (+12)
US County	2	3,234	12,717	−74.56%	+1.77	39 (+12)
wikiCS	10	11,701	216,123	−84.88%	+2.03	7 (+11)

Figure: from Table 1. Performance is reported wrt SOTA GNN.

Accuracy vs Number of Parameters

Higher accuracy with less parameters on one of the datasets (and training is also significantly faster)

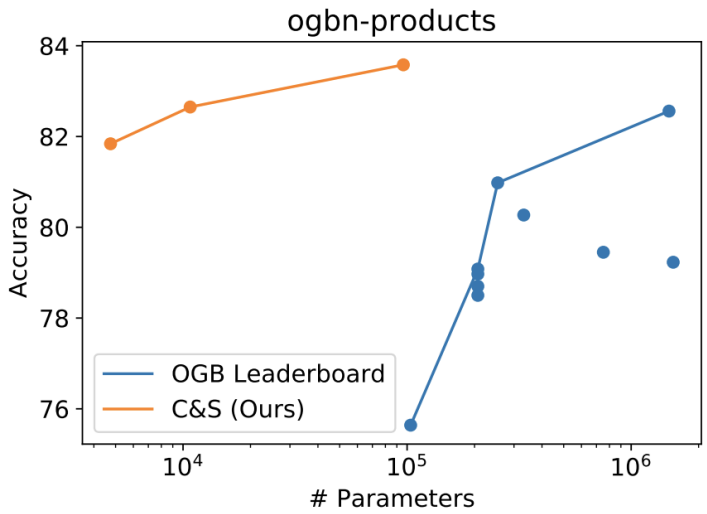
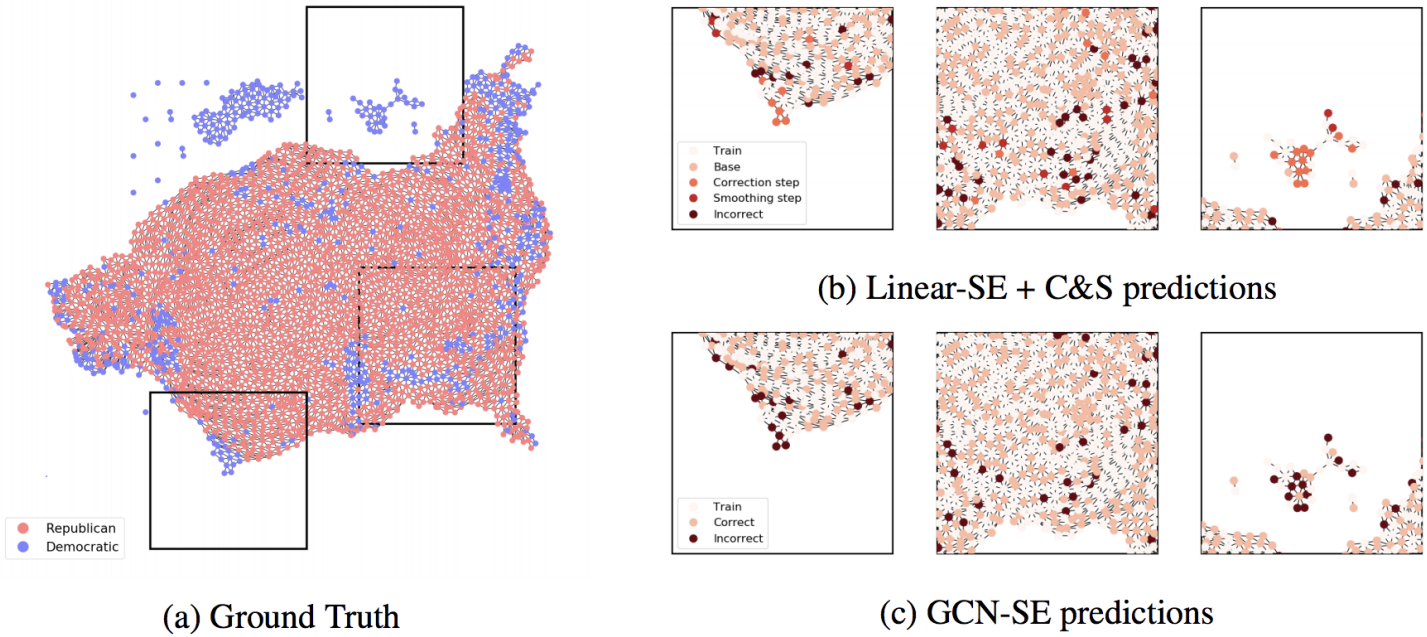


Figure: from Fig. 2

Visualizing which correction step fixed error



(a) Ground Truth

(b) Linear-SE + C&S predictions

(c) GCN-SE predictions

Figure: from Fig. 3. Colors in the correct and smooth panel show at which step labels became correct.

Summary

- Simple LP, diffusion, and GNN are fundamentally related
- Augmenting graph information with attributes, spectral features, etc. can be helpful for classifying nodes
- A base prediction is corrected according to smoothing over residual errors and encouraging closely connected nodes to have similar labels.

Recent Example Using GNNs to Study Spatial Context

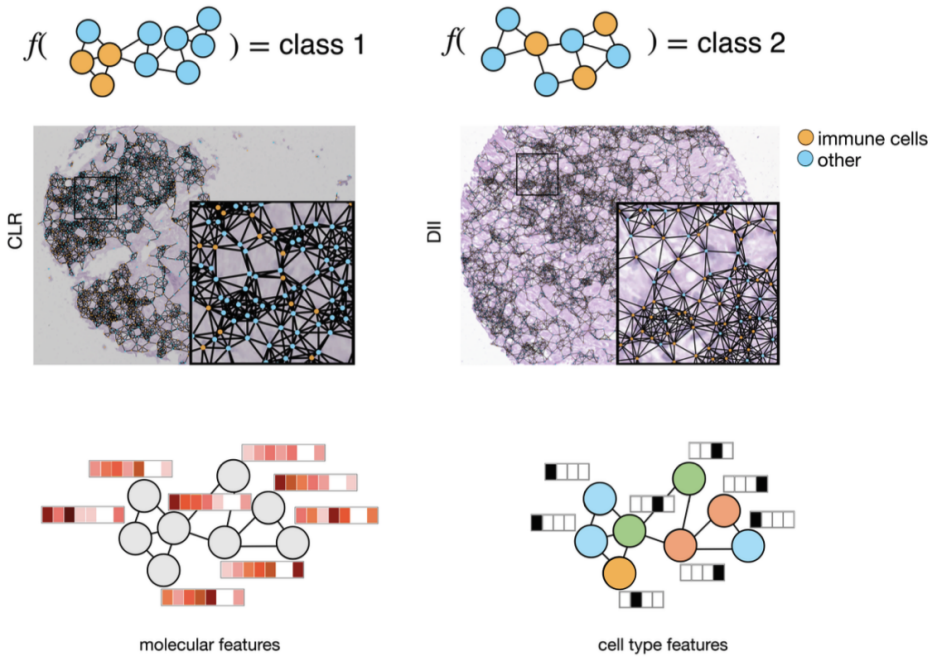


Figure: from `https://www.biorxiv.org/content/10.1101/2022.12.08.519537v1.full.pdf`. Two colorectal tumor cases Crohn's-like reaction (CLR) and diffuse inflammatory infiltration (DII) cannot be distinguished based on the spatial distribution of immune cells, but instead needs 'cellular niches'.

Ultimate Task - Sample-Level Encodings

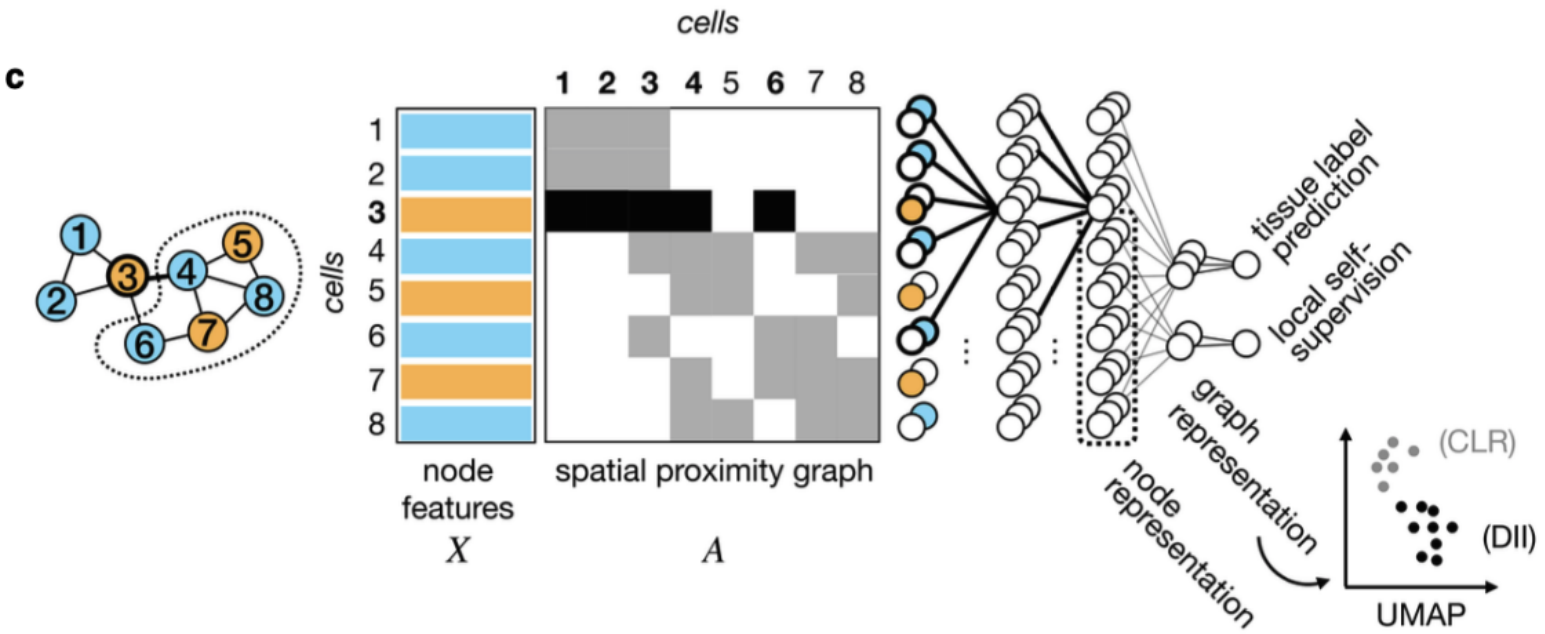


Figure: from Fischer *et al.* 2023. The learned encodings for cell's within a sample can ultimately be averaged to create a pooled feature vector that can separate tumor types.

Step 1 : Spatial Proximity Graph

Define an adjacency matrix, $\mathbf{A}$ such that with $a_{ij} = 1$ if,

$$||z_i - z_j||_2 < r$$

- z_i is the 2-D location for pixel i
- r is some user-defined radius

Graph Convolutional Network (GCN)

The node embedding layers for the GCN are defined as,

$$\mathbf{H}^{l+1} = \sigma(\mathbf{A}^*\mathbf{H}^l\mathbf{W}^l)$$

- $\mathbf{A}^* = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$
- $\mathbf{A}$ is the raw adjacency matrix and $\mathbf{D}$ is the diagonal degree matrix.
- $\mathbf{H}^l$ is the input matrix of nodes $\times$ input features
- $\mathbf{W}^l$ is a weight matrix of input features $\times$ output features

Results- Breast Cancer vs Colorectal Cancer

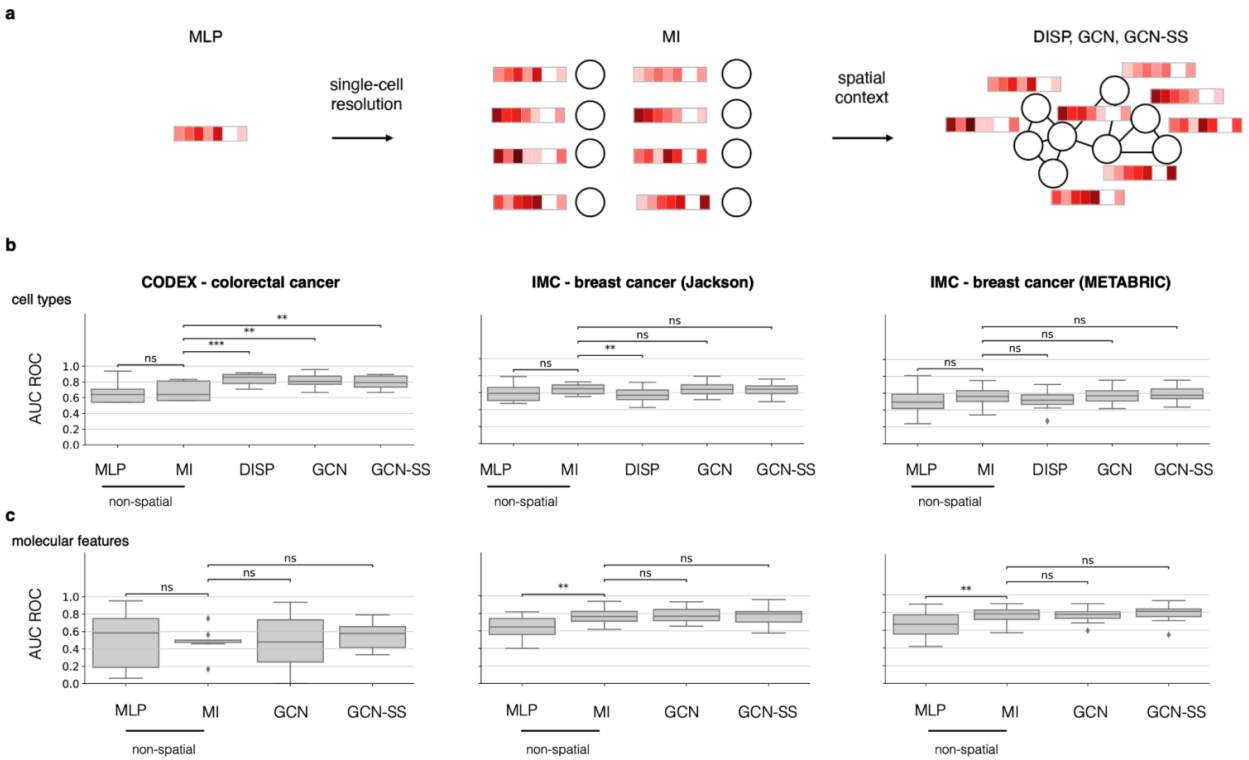


Figure: from Fischer *et al.* 2023. Results compared to just using original features for prediction. Spatial context helps things in the colorectal cancer dataset, but not so much in the breast cancer dataset.

Classical Omics Integration Problem

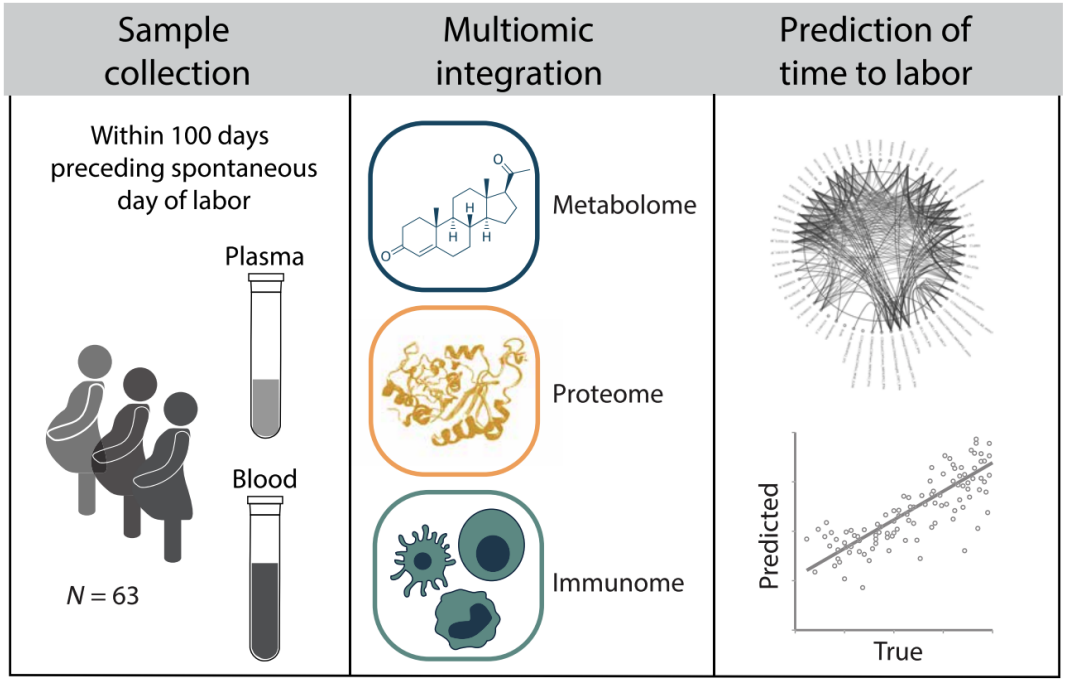


Figure: Figure from Stelzer *et al.* Science Translational Medicine. 2021. How do we leverage disparate modalities to predict something about patients, given inherent properties and quirks of each dataset?

The Cancer Genome Atlas (TCGA)

The cancer Genome Atlas was one of the first major profiling efforts, collecting diverse types of data across many patients, cancers, and biological modalities.

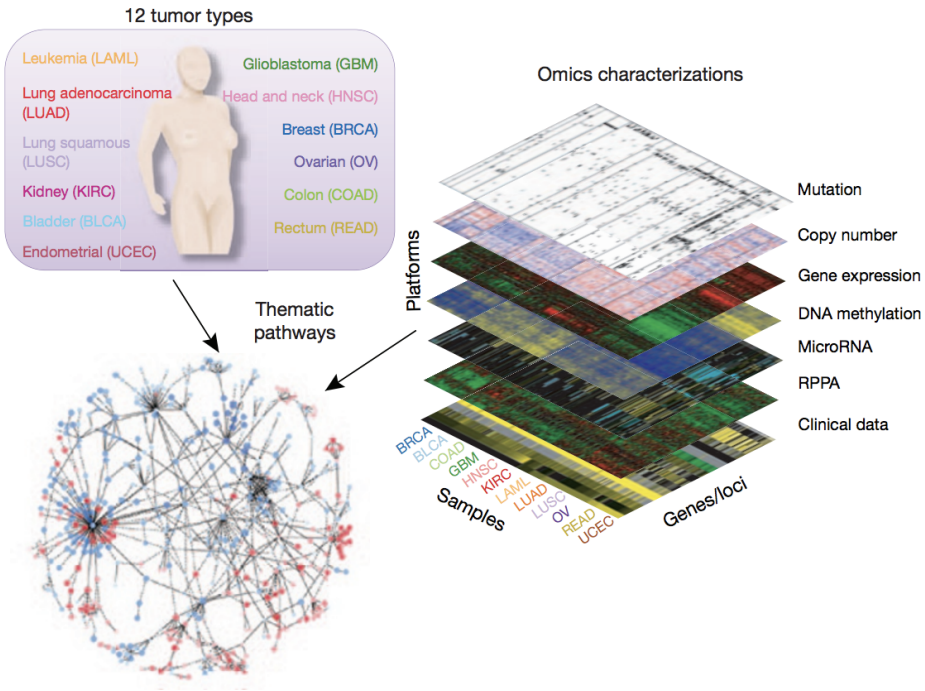


Figure: from TCGA, Nature Genetics. 2013.

FYI: LinkedOmics for Ready-to-use data with minimal pre-processing

- Download TCGA data here across many different cancers
- `http://www.linkedomics.org/login.php`

LinkedOmics "OMICS" Datatype

- **Clinical Data :** It inculdes attributes like age, overall survival, phological stage (I, II, III, IV), TNM staging, Clinical subtype, Molecular Subtype, number of lymph nodes, radiation therapy.
- **Copy Number (Level: Focal, Gene) :** Normalized copy number (SNPs) ans Copy number alterations for aggregated/segmented regions, per sample
- **miRNA (Leve: Gene, Isoform) :** Normalized signals per probe or probe set for each participant's tumor sample
- **Mutation (Level: Site, Gene) :** Mutation calls for each participant
- **Methylation (Level: Gene) :** Calculated beta values mapped to genome, per sample
- **RNAseq (Level: Gene) :** The normalized expression signal of individual Gene (transcripts), per sample
- **RPPA (Level: Analyte, Gene) :** Normalized protein expression for each gene, per sample
- **Proteomics (Level: Gene) :** Average log-ratio of sample reporter-ion to common reference of peptide ions associated with the gene in acquisitions from a specific biological Sample (Unshared Log Ratio-Average log-ratio of sample reporter-ion to common reference of peptide ions of unshared peptides only associated with the gene in acquisitions from a specific biological sample).
- **Phospho-Proteomics (Level: Site) :** Average log-ratio of sample reporter-ion to common reference of peptide ions associated with phosphorylated site combinations in acquisitions from a specific biological sample (CDAP Protein Report).
- **Glyco-Proteomics (Level: Site) :** Average log-ratio of sample reporter-ion to common reference of peptide ions associated with deglycosylated N-glycosylation site combinations in acquisitions from a specific biological sample (CDAP Protein Report).

For more information (Click here)

LinkedOmics Data Source

Cancer Type	Cohort Source	Cancer ID	Samples	Death Events	Median OS (yrs)	Permission	Link	Data Download
Adrenocortical carcinoma	TCGA	ACC	92	33	NA	Y	TCGA, GDAC	Download
Bladder urothelial carcinoma	TCGA	BLCA	412	178	2.84	Y	TCGA, GDAC	Download
Breast invasive carcinoma	TCGA	BRCA	1097	151	10.81	Y	TCGA, GDAC, CPTAC	Download
Cervical and endocervical cancers	TCGA	CESC	307	71	8.48	Y	TCGA, GDAC	Download

The Problem Also Comes up for Single-Cell

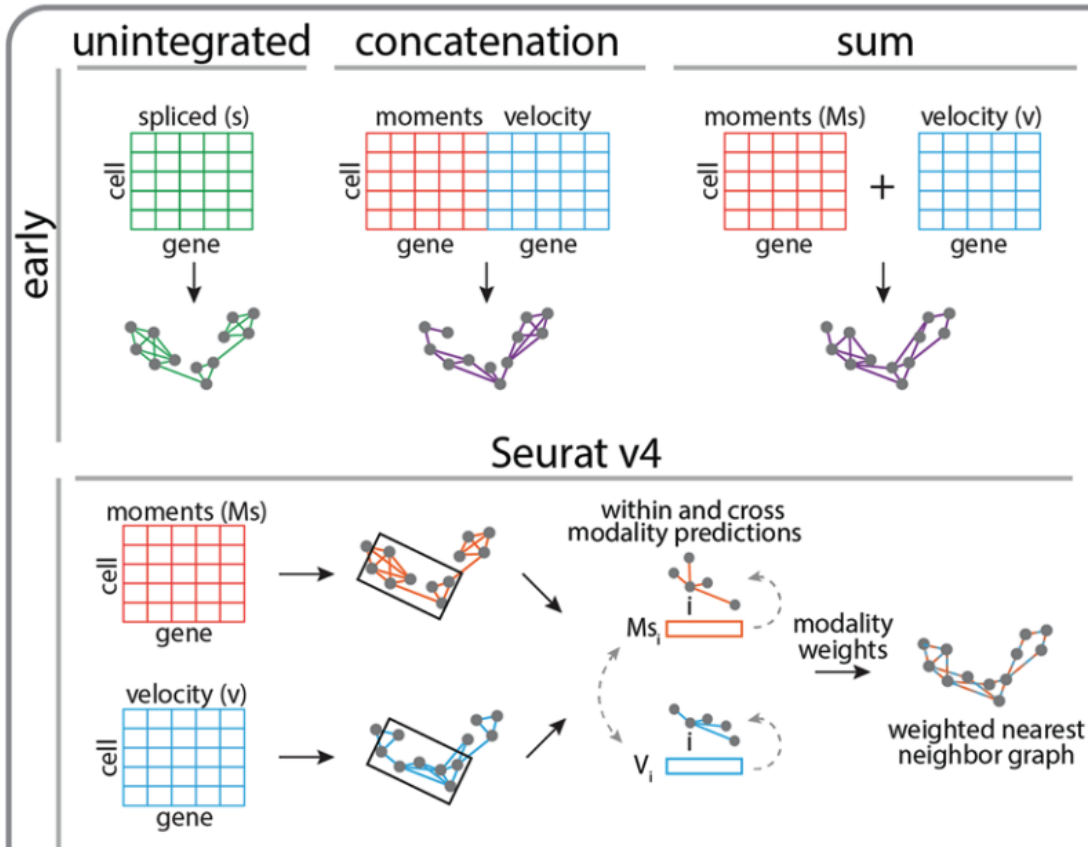


Figure: from Ranek *et al.* Genome Biology. 2022. How do we best combine various single-cell measurements to (for example) predict the label of the sample?