Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatial-spectral hypergraph attention for unsupervised segmentation of lung cancer MALDI-FTICR imaging: a leave-one-patient-out benchmark against classical clustering

Amine Bougarne

amine.bougarne@um5r.ac.ma

Mohammed V University

Research Article

Keywords: mass spectrometry imaging, MALDI-FTICR, lung cancer, hypergraph neural networks, unsupervised segmentation, leave-one-patient-out validation, METASPACE

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11327285/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Graph neural networks are increasingly used to segment mass spectrometry imaging (MSI) data without labels, but they are seldom tested on patients they have never seen, and seldom against plain clustering. I built a spatial-spectral hypergraph attention network (HyperGAT) for lung cancer MALDI-FTICR imaging and asked whether linking spatially distant regions that share a metabolic profile helps segment new patients.

Methods

Seven public lung cancer MALDI-FTICR sections were downloaded from METASPACE through its Python API and divided into 2,936 superpixels described by 108 *m/z* features. Each section became a hypergraph with spatial, non-local spectral and metabolic-module hyperedges. A two-level attention encoder was trained with a reconstruction loss and a soft *k*-means loss. HyperGAT was compared with seven baselines and six ablated variants by leave-one-patient-out cross-validation (LOPO-CV, three seeds), with the number of clusters chosen inside each training fold.

Results

Every fold selected two clusters. HyperGAT gave the best silhouette among the learned models (0.297 ± 0.047, against 0.161 to 0.217 for graph attention networks and a Gaussian mixture) and was on par with a 3D convolutional autoencoder. Removing any of its components lowered performance, by 6% to 49%. Yet *k*-means on standardised spectra beat HyperGAT in all seven held-out patients (silhouette 0.445 ± 0.074; Davies-Bouldin 1.28 vs 1.80; spatial coherence 0.954 vs 0.930). None of these differences remained significant after Holm correction. A distance-based margin score rose at the boundaries between metabolic compartments in every held-out patient (mean AUC 0.684).

Conclusions

On this small, low-resolution cohort, hypergraph attention was the strongest graph model I tested, but it did not outperform classical clustering. Deep unsupervised MSI methods should be reported with patient-level validation, per-section normalisation and a *k*-means baseline. The code and data retrieval scripts are public.

1. Introduction

Lung cancer causes more cancer deaths than any other tumour type [1]. When the disease is resectable, surgery offers the best chance of cure, and the molecular make-up of the tumour and of the tissue

around it matters both to pathologists and to researchers. Mass spectrometry imaging (MSI) measures hundreds of lipids and metabolites directly in a tissue section. Coupled with Fourier-transform ion cyclotron resonance (FTICR), matrix-assisted laser desorption/ionisation (MALDI) reaches the mass resolving power needed to separate isobaric species [2]. METASPACE now hosts thousands of public datasets with false-discovery-rate (FDR) controlled annotations [3], which makes reproducible computational work possible without access to an instrument.

Histological annotation of MSI data is rarely available, so the usual first analysis is unsupervised segmentation [4, 5]. Most pipelines reduce dimensionality with PCA or UMAP [6] and then cluster with *k*-means or a Gaussian mixture. Convolutional autoencoders and graph neural networks (GNNs) have been proposed to add spatial context. A pairwise graph, however, can only connect two nodes at a time, usually neighbours in the tissue. Metabolic programmes do not follow that rule: the same lipid profile may appear in several separate regions of a section. A hypergraph handles this naturally, because one hyperedge can join any number of nodes [7–9].

The published evidence has two weak points. Models are often trained and tested on pixels pooled from the same sections. Information about each patient then leaks into the test set and inflates the scores. And deep models are often compared only with other deep models, so we do not know whether they beat ordinary clustering at all.

In this study I describe HyperGAT, a spatial-spectral hypergraph attention network for unsupervised MALDI-FTICR segmentation, and test it under conditions designed to be hard to game. The data are retrieved through the METASPACE API, so anyone can rerun the analysis. The hypergraph mixes spatial, non-local spectral and metabolic-module hyperedges, and an *m/z* gate gives a handle on which ions the model relies on. The benchmark uses leave-one-patient-out validation, chooses the number of clusters without touching the test patient, and pits HyperGAT against seven baselines and six ablations with paired non-parametric tests. I report where the hypergraph helped and where it did not. Figure 1 outlines the workflow.

2. Materials and methods

2.1 Data

I analysed seven public human lung cancer sections from the METASPACE "clinic" series (Table 1). According to the repository metadata, they were acquired by MALDI-FTICR in positive ion mode, with a resolving power of 500,000 at *m/z* 184 and a pixel size of 100 µm. The datasets were found and downloaded with the official *metaspace2020* Python client. For each one I retrieved annotations and monoisotopic ion images up to 50% FDR from every database available on the dataset (SwissLipids, HMDB v4, HMDB-endogenous v4 and CoreMetabolome v3). Annotations were merged by molecular formula and adduct, keeping the one with the lowest FDR. I also downloaded the total ion current (TIC) image computed over the full spectrum. No pathology annotation exists for these sections.

Table 1
The seven MALDI-FTICR lung cancer sections retrieved from METASPACE.

Patient	METASPACE ID	Image (px)	Tissue (px)	Superpixels	Ions at FDR ≤ 10%*	Features not detected†
001	2026-09-21_04h45m20s	47 × 117	3,670	447	16	7
002	2026-09-21_04h54m54s	49 × 66	2,017	246	48	11
004	2026-09-21_05h42m31s	69 × 94	4,555	559	56	5
009 – 1	2026-09-30_07h26m20s	31 × 55	910	111	38	8
018	2026-09-29_05h35m31s	82 × 176	10,015	1,228	21	9
019	2026-09-30_07h27m09s	39 × 47	1,302	160	17	14
021 – 1	2026-09-30_07h27m56s	44 × 49	1,536	185	16	11
Total			**24,005**	**2,936**		

* Unique formula and adduct pairs annotated at FDR ≤ 10%, all databases combined. † Number of the 108 shared features that were not detected (FDR ≤ 50%) in the section and were set to zero.

2.2 Shared feature space and preprocessing

Testing on unseen patients requires every patient to be described by the same features. I kept an ion when it was annotated at FDR ≤ 10% in at least one section and detected at FDR ≤ 50% in at least 70% of the sections. This left 108 features. Most of them were annotated with confidence in a single section (68), 29 in two sections and 11 in three or more. Features that were not detected in a section were set to zero.

Tissue was separated from background by Otsu thresholding of the log TIC image, followed by morphological opening and hole filling. Spectra were divided by their TIC and rescaled to the median TIC, clipped at the 99th percentile of each ion to remove hotspots, and log-transformed. I then ran simple linear iterative clustering (SLIC) [10] on the first three principal components of each section, computed on that section alone. Superpixels held about 8 pixels (0.08 mm²), which gave between 111 and 1,228 superpixels per section (Table 1, Fig. 2). Each superpixel was described by its mean spectrum.

Finally, every feature was centred and scaled within each section, using only that section's superpixels. The step uses no label and is applied in the same way to the held-out section. I added it after early runs without it often put every superpixel of the held-out patient into a single cluster. Those early runs also

differed in other settings, so I mention this as an observation from development and not as a controlled result.

2.3 Hypergraph construction

Each section was turned into a hypergraph $G = (V, E, W)$ whose nodes are superpixels (Fig. 3). Spatial hyperedges join a superpixel to its 8 nearest neighbours in the tissue, with weight 1. Spectral hyperedges join a superpixel to its 10 nearest neighbours in cosine distance between spectra, wherever they lie in the section; their weight is the mean cosine similarity. Module hyperedges come from a *k*-means partition of the spectra, with roughly one module per 25 superpixels; each module is one hyperedge, weighted by its mean cohesion. A hyperedge holds about 10 superpixels on average (34 to 54 at most) and every node belongs to 21 hyperedges.

2.4 Model and training

The input spectra are first multiplied by an *m/z* gate $\gamma = F \cdot \mathrm{softmax}(g)$, with g learned. The gate has a fixed budget (its mean is 1), and a penalty on its normalised entropy pushes it to favour some ions over others. Each HyperGAT layer [8, 9] then works in two steps on the projected features h. Nodes are pooled into hyperedges with attention weights $\alpha(v,e)$ = softmax over $v \in e$ of $\mathrm{LeakyReLU}(a_1 \cdot h_v)$, giving $f_e = \Sigma\, \alpha\, h_v$. Hyperedges are then pooled back into nodes with weights $\beta(e,v)$ = softmax over $e \ni v$ of $\mathrm{LeakyReLU}(a_2 \cdot [h_v \,\|\, f_e]) + \log w_e$. The encoder has two such layers: 128 units (4 heads of 32, ELU, dropout 0.2), then a 64-dimensional embedding. A two-layer decoder reconstructs the input.

The loss adds the mean squared reconstruction error to a soft *k*-means term on the normalised embeddings, with assignments $q = \mathrm{softmax}(-d/\tau)$ and $\tau = 0.1$ [11]. To avoid empty clusters without forcing clusters of equal size, a hinge penalty applies only when a cluster receives less than 5% of the assignments. Training runs 150 epochs of reconstruction alone, initialises the centroids by *k*-means, then runs 150 epochs on the full loss with Adam [12] (learning rate 10^{-3}, weight decay 5×10^{-4}). Everything was written in PyTorch [13].

2.5 Baselines and ablations

All baselines received the same inputs: *k*-means on the standardised spectra; PCA with 20 components followed by *k*-means; UMAP with 10 components followed by *k*-means [6]; a Gaussian mixture on the PCA components; a 3D convolutional autoencoder fed with 7 × 7-pixel spectral patches centred on each superpixel; and two graph attention networks (GAT) [14] trained with the HyperGAT loss, one on spatial pairwise edges and one on spatial plus spectral pairwise edges. The six ablations each removed one part of HyperGAT: the spectral and module hyperedges, the spatial hyperedges, the attention (replaced by degree-normalised averaging, which amounts to an HGNN [7]), the *m/z* gate, the reconstruction loss, or the clustering loss (in which case *k*-means was run on the embeddings afterwards).

2.6 Validation and metrics

In each of the seven folds, everything that is fitted (scaler, network weights, centroids, PCA, UMAP, mixture model) was estimated on six patients and then applied to the seventh. The number of clusters K was picked from {2, 3, 4} within each fold by maximising the mean silhouette of k-means on the six training patients, and all methods used that K in the fold. Each configuration ran with three seeds. I also built a consensus partition for every method from the co-association matrix of its three runs, using average-linkage clustering [15].

With no ground truth, I compared methods on intrinsic criteria computed in one evaluation space for all of them, namely the spectra standardised with the fold scaler. These were the silhouette coefficient [16] (higher is better), the Davies-Bouldin index [17] (lower is better), the Calinski-Harabasz index [18] (higher is better) and a spatial coherence score (SCS), the share of 4-adjacent tissue pixel pairs that receive the same label. I also counted runs in which a method assigned the whole held-out section to one cluster. Patient-level values were averaged over seeds and compared with two-sided Wilcoxon signed-rank tests paired by patient, with Holm correction within each metric [19]. With seven patients the smallest possible two-sided p is 0.0156.

2.7 Margin score and candidate ions

For every superpixel I computed a margin score, the ratio of its latent distance to the nearest centroid over its distance to the second nearest. A value close to 1 means the superpixel sits between two metabolic compartments. I compared border superpixels (those with a neighbour in another cluster) with interior ones using one-sided Mann-Whitney tests, reported the AUC and applied Benjamini-Hochberg correction [20]. Ions were ranked by the mean of three ranks: gate weight, integrated-gradient attribution [21] and differential effect size between clusters. Differential tests used only the sections in which the ion was detected, and were run both on superpixels and on patient means (paired Wilcoxon test), the latter to avoid pseudo-replication.

3. Results

3.1 Cohort

The sections were small, from 910 to 10,015 tissue pixels, and produced 2,936 superpixels in total (Table 1, Fig. 2). Annotation was thin. Only 16 to 56 unique ions per section passed FDR ≤ 10%, and most of the shared features were annotated with confidence in one section only. The hypergraphs contained between 2,331 and 25,788 node-hyperedge incidences.

3.2 Benchmark under leave-one-patient-out validation

All seven folds chose two clusters (in the first fold, the training silhouette was 0.507 for $K = 2$, 0.391 for $K = 3$ and 0.368 for $K = 4$). No method collapsed to a single cluster on a held-out patient. Table 2 and Fig. 4 give the results.

HyperGAT had the highest silhouette of the learned models, 0.297 ± 0.047, and the lowest Davies-Bouldin index, 1.80 ± 0.43. It did better than the spatial GAT (0.161 ± 0.069) in all seven patients, better than the spatial-spectral GAT (0.217 ± 0.057) and the Gaussian mixture (0.187 ± 0.225) in six, and better than UMAP followed by *k*-means (− 0.018 ± 0.105) in all seven. Against the 3D convolutional autoencoder (0.265 ± 0.061) it won in four patients out of seven, so the two are effectively tied.

The picture changes once plain clustering enters the comparison. *k*-means and PCA followed by *k*-means beat HyperGAT in every held-out patient, on silhouette (0.445 ± 0.074 and 0.446 ± 0.077), on Davies-Bouldin (1.28 against 1.80) and on spatial coherence (0.954 to 0.955 against 0.930). Figure 5 shows how consistent this was from one patient to the next. The uncorrected *p*-values were 0.0156, the floor for seven patients, and none survived Holm correction ($p = 0.109$; Table 3). Consensus over seeds raised HyperGAT's silhouette to 0.329 ± 0.078 without changing the order (PCA followed by *k*-means, 0.447 ± 0.076; Holm $p = 0.125$). On the segmentation maps (Supplementary Fig. S1), *k*-means and HyperGAT find a similar two-compartment layout, and HyperGAT draws more ragged boundaries.

Table 2
Segmentation performance under LOPO-CV, as mean ± s.d. over the seven held-out patients (each patient value is the mean of three seeds). The best value in each column is in bold.

Method	**Silhouette ↑**	**Davies-Bouldin ↓**	**SCS ↑**
k-means	0.445 ± 0.074	1.284 ± 0.312	0.954 ± 0.021
PCA + *k*-means	0.446 ± 0.077	1.282 ± 0.306	0.955 ± 0.021
UMAP + *k*-means	−0.018 ± 0.105	4.127 ± 1.052	0.903 ± 0.038
Gaussian mixture	0.187 ± 0.225	2.690 ± 0.766	0.898 ± 0.056
3D-CNN autoencoder	0.265 ± 0.061	2.001 ± 0.547	0.951 ± 0.015
GAT (spatial)	0.161 ± 0.069	2.780 ± 0.359	0.948 ± 0.014
GAT (spatial + spectral)	0.217 ± 0.057	2.081 ± 0.346	0.919 ± 0.023
HyperGAT	0.297 ± 0.047	1.801 ± 0.429	0.930 ± 0.018

SCS, spatial coherence score. Calinski-Harabasz index: HyperGAT 72.6 ± 31.1, *k*-means 94.1 ± 92.3. No method collapsed to a single cluster on any held-out patient.

Table 3
HyperGAT against each baseline: number of patients in which HyperGAT did better, two-sided Wilcoxon signed-rank *p* (paired by patient) and Holm-corrected *p* within each metric.

Comparator	Silhouette	Davies-Bouldin	SCS
k-means	0/7; 0.016; 0.109	0/7; 0.016; 0.109	0/7; 0.016; 0.109
PCA + *k*-means	0/7; 0.016; 0.109	0/7; 0.016; 0.109	0/7; 0.016; 0.109
UMAP + *k*-means	7/7; 0.016; 0.109	7/7; 0.016; 0.109	5/7; 0.156; 0.469
Gaussian mixture	6/7; 0.219; 0.438	6/7; 0.047; 0.141	5/7; 0.156; 0.469
3D-CNN autoencoder	4/7; 0.578; 0.578	5/7; 0.156; 0.313	1/7; 0.047; 0.188
GAT (spatial)	7/7; 0.016; 0.109	7/7; 0.016; 0.109	0/7; 0.016; 0.109
GAT (spatial + spectral)	6/7; 0.047; 0.141	5/7; 0.297; 0.313	5/7; 0.156; 0.469

3.3 Ablation

Each part of HyperGAT earned its place (Table 4, Fig. 6). Without the spectral and module hyperedges, the silhouette fell by 44% and the Davies-Bouldin index rose to 2.81. This variant also had the highest spatial coherence (0.958), which suggests that purely spatial message passing smooths the maps at the expense of metabolic contrast. The largest losses came from dropping the reconstruction loss (− 49%) or the clustering loss (− 37%). Attention (− 17%) and the *m/z* gate (− 6%) mattered less. The variant without the reconstruction loss was also the only one that collapsed to a single cluster, in one run out of 21.

Table 4
Ablation under LOPO-CV (mean ± s.d. over held-out patients).

Variant	Silhouette ↑	Change	Davies-Bouldin ↓	SCS ↑
HyperGAT (full model)	0.297 ± 0.047		1.801	0.930
No spectral or module hyperedges	0.167 ± 0.041	−43.8%	2.805	0.958
No spatial hyperedges	0.240 ± 0.058	−19.2%	2.001	0.906
No attention (HGNN)	0.248 ± 0.041	−16.5%	1.972	0.946
No *m/z* gate	0.278 ± 0.069	−6.4%	1.721	0.934
No reconstruction loss	0.151 ± 0.062	−49.2%	2.203	0.943
No clustering loss	0.186 ± 0.023	−37.4%	2.239	0.914

The variant without reconstruction loss collapsed to one cluster in one run (patient 001, seed 2); its silhouette was set to 0 for that run and its Davies-Bouldin value was left out.

3.4 Sensitivity analyses

Cluster quality went down as *K* went up on the model trained on all patients: the silhouette was 0.217, 0.211, 0.178, 0.078 and 0.051 for *K* = 2 to 6, and spatial coherence fell from 0.925 to 0.801 (Table 5, Fig. 7). This agrees with the two clusters chosen inside the training folds. The neighbourhood sizes had a moderate effect, and the defaults (8 spatial and 10 spectral neighbours) gave the best silhouette in both sweeps.

In the embedding space, clusters were well separated (silhouette 0.711) and patients were well mixed (patient silhouette − 0.167), which indicates that per-section scaling removed most of the patient-specific signal. Attention, on the other hand, barely discriminated between hyperedge types: the ratio of observed to uniform attention was 0.94 for spatial, 1.02 for spectral and 0.98 for module hyperedges. One HyperGAT fold took 15.2 s on a GPU with 45,272 parameters; *k*-means took 0.14 s.

Table 5
Sensitivity analyses (mean ± s.d. over patients). Top: number of clusters *K* on the model trained on all patients. Bottom: neighbourhood sizes under LOPO-CV (seed 0).

Setting	Silhouette ↑	Davies-Bouldin ↓	SCS ↑
K = 2	0.217 ± 0.029	1.772 ± 0.183	0.925 ± 0.025
K = 3	0.211 ± 0.043	1.670 ± 0.251	0.882 ± 0.036
K = 4	0.178 ± 0.059	1.574 ± 0.317	0.861 ± 0.032
K = 5	0.078 ± 0.050	2.144 ± 0.508	0.818 ± 0.032
K = 6	0.051 ± 0.046	2.447 ± 0.587	0.801 ± 0.034
4 spatial neighbours	0.266 ± 0.096	1.931 ± 0.554	0.938 ± 0.022
8 spatial neighbours (default)	0.300 ± 0.060	1.811 ± 0.378	0.937 ± 0.023
12 spatial neighbours	0.244 ± 0.075	1.832 ± 0.384	0.918 ± 0.025
5 spectral neighbours	0.283 ± 0.089	1.696 ± 0.549	0.933 ± 0.021
10 spectral neighbours (default)	0.316 ± 0.090	1.709 ± 0.571	0.936 ± 0.025
20 spectral neighbours	0.264 ± 0.072	1.939 ± 0.409	0.938 ± 0.023

3.5 Margin score and candidate ions

The margin score was higher at compartment borders than inside compartments in all seven held-out patients (mean AUC 0.684), and the difference was significant in six after Benjamini-Hochberg correction (Supplementary Fig. S2). Without histology, these borders are transitions between metabolic compartments; I cannot equate them with tumour margins.

No candidate ion reached significance at the patient level, the smallest *q* being 0.096 (Table 6; Supplementary Fig. S3). The gate stayed close to uniform (γ between 0.86 and 1.30), and its ranking of ions was only moderately stable across folds (top-20 Jaccard index 0.39, Spearman ρ = 0.69). The strongest signals came from an ion at *m/z* 735.468, richer in compartment C1 (log2 fold change 2.90, AUC 0.81, same direction in six of six patients), an ion at *m/z* 758.546, poorer in C2 (− 3.76, six of six), whose annotation hesitates between ether-linked PC and PG species, and a possible PE plasmalogen at *m/z* 723.520. All these annotations rest on MS1 accurate mass and need tandem MS to be confirmed.

Table 6
Top candidate ions. Annotations are putative (MS1 accurate mass, METASPACE FDR control). Differential statistics use only the sections in which the ion was detected.

m/z	Formula and adduct	Putative annotation	Cluster	log2FC	AUC	Same sign	q (patient)
758.546	C40H82NO7P + K or C42H79O9P (M+)	PC(O-32:0) or PG(O-36:3) isomers	C2	−3.76	0.17	6/6	0.096
723.520	C41H74NO7P (M+)	PE(P-36:4) isomers	C1	−2.91	0.20	5/7	0.156
241.180	C14H24O3 + H	unassigned	C1	−1.14	0.29	6/7	0.096
735.468	C41H66O11 + H	unassigned	C1	2.90	0.81	6/6	0.096

The first two features in the automated ranking were the same peak (*m/z* 758.546) annotated with two candidate formulas; they are merged here. Matches to general databases that are not plausible lipids (a flavour compound for *m/z* 241.180 and a plant natural product for *m/z* 735.468) are listed as unassigned.

4. Discussion

I set out to test whether hypergraph attention, by linking distant regions with similar metabolic profiles, would segment unseen lung cancer sections better than existing methods. The answer has two halves. Among learned models, the hypergraph clearly helped. HyperGAT beat pairwise graph attention networks in almost every patient, and the ablation points to the non-local spectral hyperedges as the main reason. Against *k*-means on standardised spectra, though, HyperGAT lost in every patient.

I see three reasons for this. The cohort is small and coarse: 2,936 superpixels at 100 µm give a model with about 45,000 parameters little to learn from, and give graph propagation little spatial structure to work with. The metabolic contrast is also weak. Every method found two compartments, and none reached a silhouette of 0.45. The third reason is a bias in the evaluation itself. Silhouette and Davies-Bouldin are computed in the input space, which is the very space *k*-means optimises, whereas HyperGAT optimises a smoothed latent space. I chose this common space so that every method would be judged by the same rule, but the choice favours *k*-means. It is less easy to explain away the spatial coherence

result: *k*-means produced the most coherent maps, so the spatial smoothing that motivated graph models did not pay off here. Only a comparison with histology can say which partition is more meaningful biologically.

One practical lesson came out of the development work. Without per-section scaling, a model trained on six patients often placed the whole seventh section in one cluster. Validation on pooled pixels would never show this failure. Intensity offsets between sections and ions missing from some sections are common in multi-section MSI, and scaling each section on its own is a cheap way to guard against them. The good mixing of patients in the embedding space supports this.

The interpretation outputs should be read as exploratory. Attention was almost uniform across hyperedge types, the gate was only mildly selective and no ion was significant at the patient level. Ether lipids and plasmalogens are plausible in lung tissue, but MS1 annotations are ambiguous (the PC or PG assignment of *m/z* 758.546 is a good example) and they need MS/MS and histology before any biological claim.

4.1 Limitations

The study rests on seven sections from one series, without pathology annotations. I therefore make no claim about tumour, stroma or surgical margins, and could not compute supervised metrics such as the adjusted Rand index. Statistical power is low: with seven patients the smallest two-sided p is 0.0156, so the lack of significance after correction does not mean the methods are equivalent. I planned an external validation on independent METASPACE datasets, but it could not be completed. The feature space was limited to annotated ions, most of them confidently identified in a single section.

4.2 Recommendations

For studies that propose deep or graph-based unsupervised methods for MSI, I would ask for patient-level validation in which every fitted component is estimated on training patients, for *k*-means on standardised spectra as a baseline, for per-section normalisation, for collapse rates and seed stability to be reported, and for comparison with histology whenever clinical relevance is claimed.

5. Conclusions

Of the graph-based approaches I tested for segmenting unseen lung cancer MALDI-FTICR sections, spatial-spectral hypergraph attention performed best, and its non-local spectral hyperedges were the component that mattered most. On a small, low-resolution cohort, it still did not beat classical clustering of standardised spectra. Larger cohorts at higher resolution, with co-registered histology, are needed to find out whether hypergraph models bring gains that matter clinically. The full pipeline, from METASPACE retrieval to figures and statistics, is public so that others can run that test.

Declarations

Competing interests:

The author declares no competing interests.

Ethics approval and consent to participate:

Not applicable. This study used only publicly available, de-identified datasets from the METASPACE repository; no new human data were collected.

Consent for publication:

Not applicable.

Software

Python 3.13, PyTorch 2.11 (CUDA 12.8), scikit-learn 1.6.1 [22], scikit-image 0.25.2, SciPy 1.16.3, umap-learn 0.5.12, metaspace2020 2.0.9.

Funding:

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author Contribution

A.B. designed the study, wrote the code, ran the analyses and wrote the manuscript.

Data Availability

All data are public on METASPACE (dataset IDs in Table 1) and were retrieved with the metaspace2020 Python client.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
2. Alexandrov T. Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. Annu Rev Biomed Data Sci. 2020;3:61–87.
3. Palmer A, Phapale P, Chernyavsky I, et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. Nat Methods. 2017;14(1):57–60.

4. Alexandrov T, Becker M, Deininger SO, et al. Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. J Proteome Res. 2010;9(12):6535–46.
5. Verbeeck N, Caprioli RM, Van de Plas R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. Mass Spectrom Rev. 2020;39(3):245–91.
6. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426. 2018.
7. Feng Y, You H, Zhang Z, Ji R, Gao Y. Hypergraph neural networks. Proc AAAI Conf Artif Intell. 2019;33:3558–65.
8. Bai S, Zhang F, Torr PHS. Hypergraph convolution and hypergraph attention. Pattern Recognit. 2021;110:107637.
9. Ding K, Wang J, Li J, Li D, Liu H. Be more with less: Hypergraph attention networks for inductive text classification. Proc EMNLP. 2020:4927–36.
10. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell. 2012;34(11):2274–82.
11. Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. Proc ICML. 2016;48:478–87.
12. Kingma DP, Ba J. Adam: A method for stochastic optimization. Proc ICLR. 2015.
13. Paszke A, Gross S, Massa F et al. PyTorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 2019;32.
14. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. Proc ICLR. 2018.
15. Fred ALN, Jain AK. Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell. 2005;27(6):835–50.
16. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
17. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979;1(2):224–7.
18. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3(1):1–27.
19. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6(2):65–70.
20. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57(1):289–300.
21. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proc ICML. 2017;70:3319–28.
22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

Figures

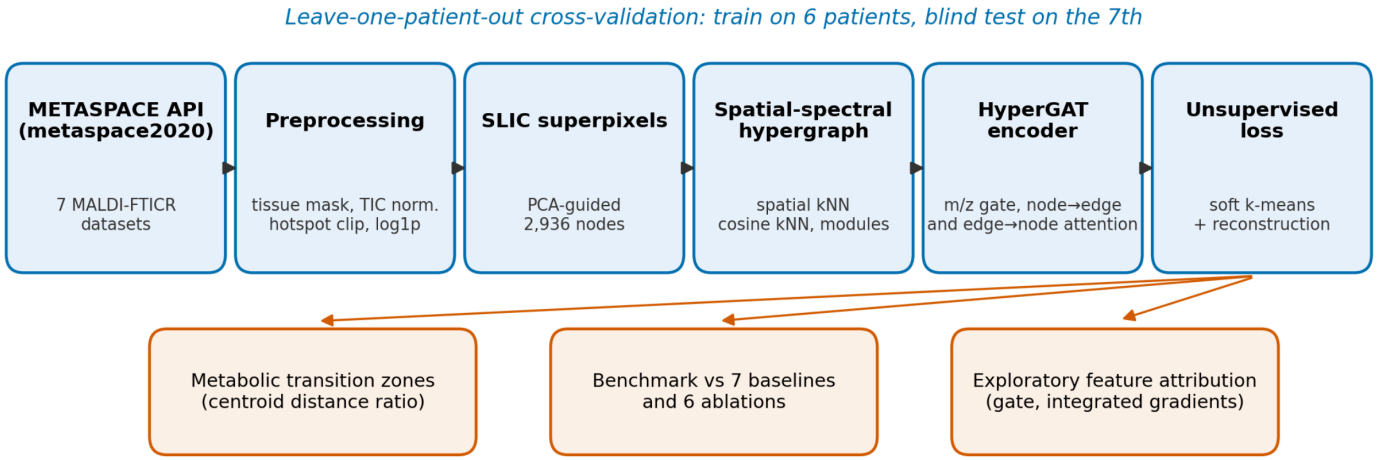


Figure 1

Analysis pipeline, from METASPACE retrieval to the downstream analyses, all carried out under leave-one-patient-out cross-validation.

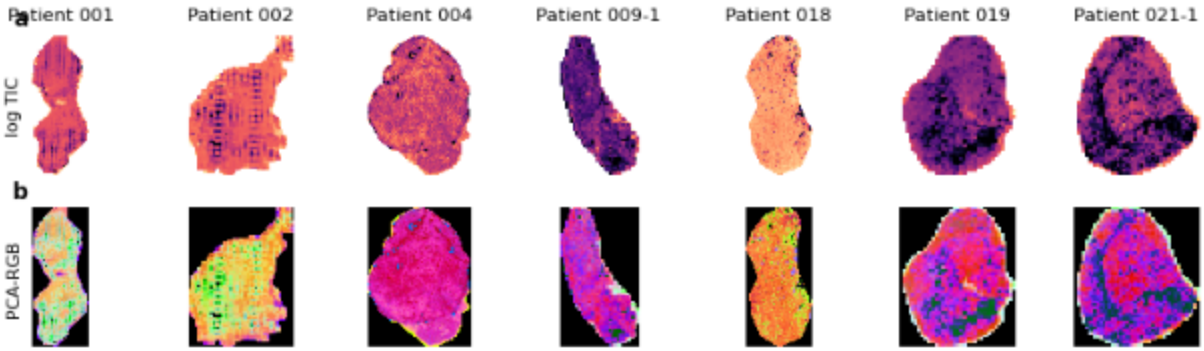


Figure 2

The seven sections. (a) Log total ion current within the tissue mask. (b) Composite of the first three principal components of the normalised ion images, shown as RGB.

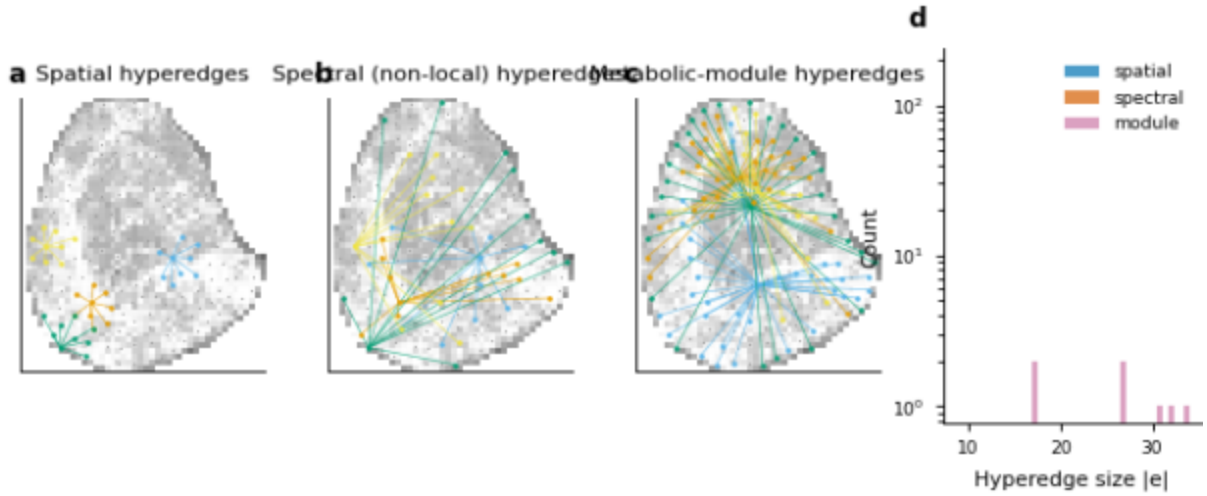


Figure 3

Hypergraph construction in one section. (a) Spatial, (b) non-local spectral and (c) metabolic-module hyperedges drawn for a few randomly chosen superpixels. (d) Distribution of hyperedge sizes.

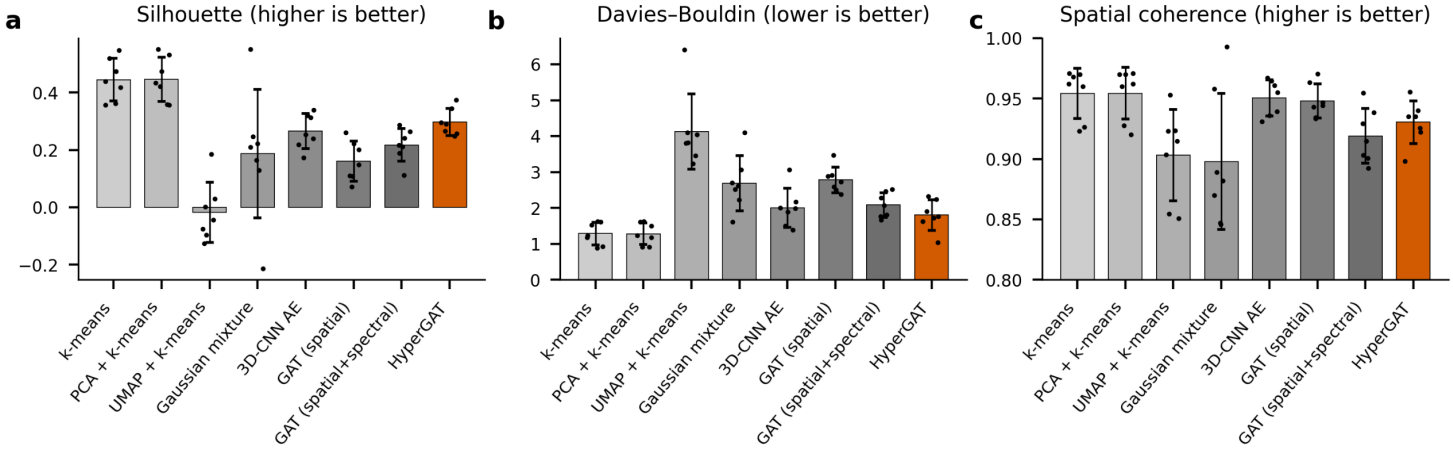


Figure 4

Benchmark under LOPO-CV. (a) Silhouette, (b) Davies-Bouldin index and (c) spatial coherence. Bars give the mean ± s.d. over the seven held-out patients and dots the individual patients. HyperGAT is shown in orange.

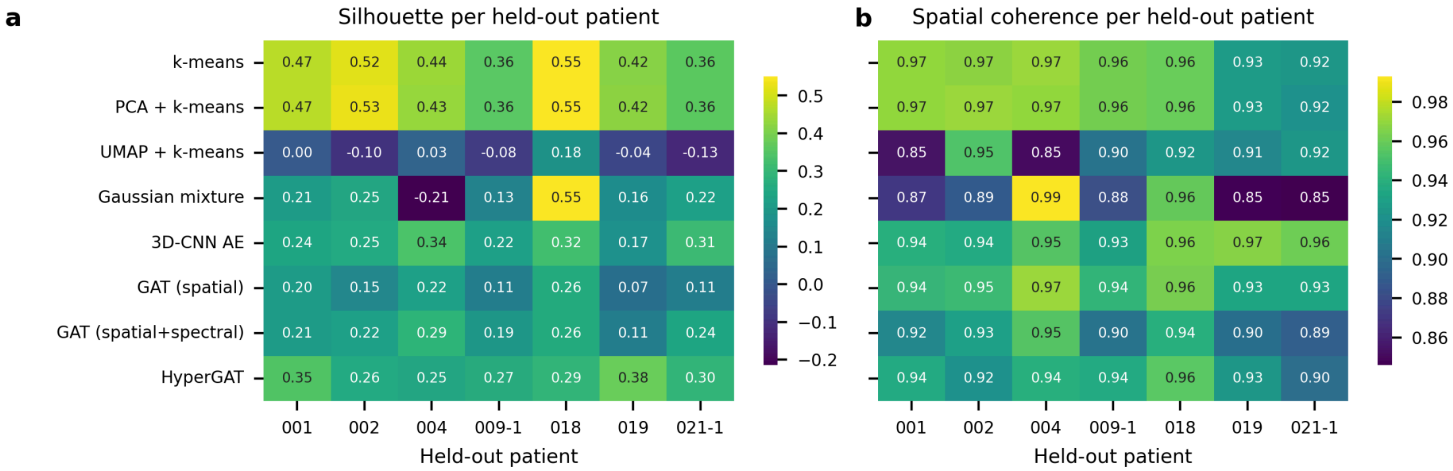


Figure 5

Results for each held-out patient. (a) Silhouette and (b) spatial coherence of every method.

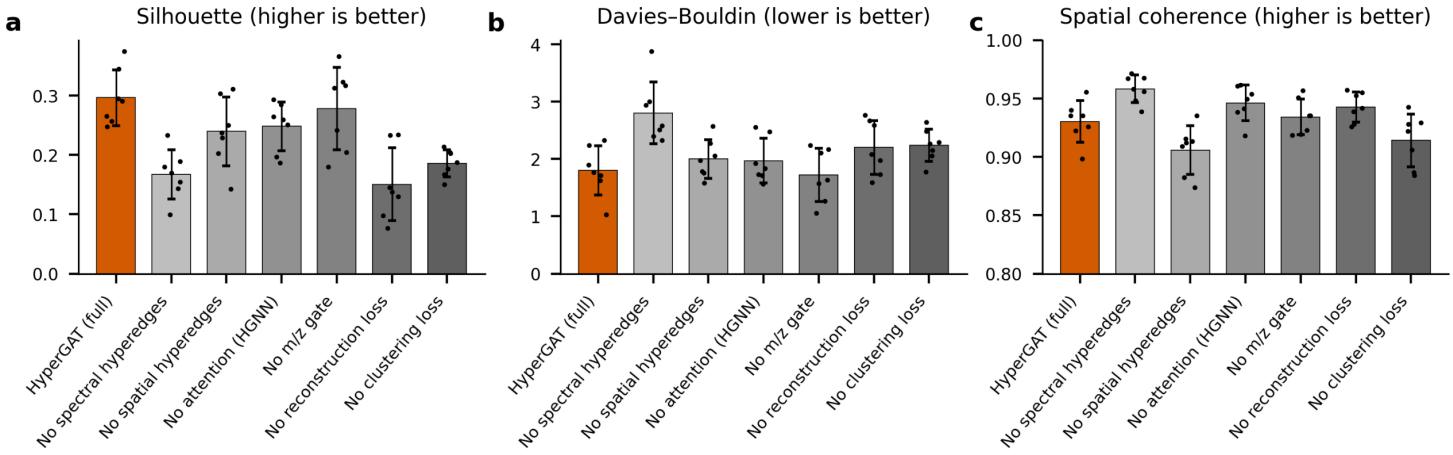


Figure 6

Ablation. (a) Silhouette, (b) Davies-Bouldin index and (c) spatial coherence of the full model and of each variant.

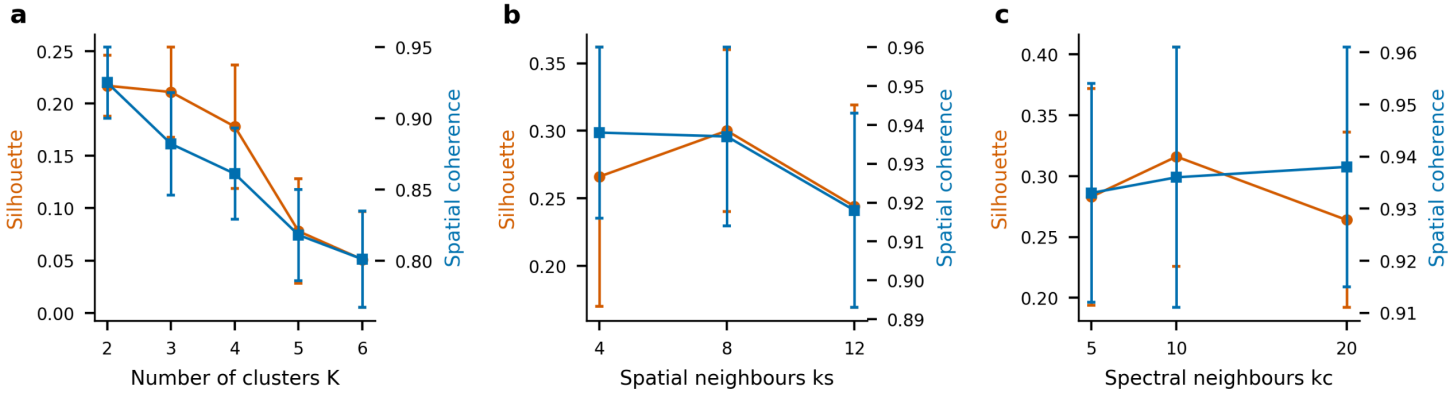


Figure 7

Sensitivity of HyperGAT to (a) the number of clusters, (b) the number of spatial neighbours and (c) the number of spectral neighbours. Orange, silhouette (left axis); blue, spatial coherence (right axis).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigS6.pdf
- SupplementaryFigS2.pdf
- SupplementaryFigS5.pdf
- SupplementaryFigS3.pdf
- SupplementaryFigS1.pdf

- SupplementaryFigS4.pdf