GRANULARITY GATING-GUIDED LABEL-SPECIFIC METRIC LEARNING FOR MULTI-LABEL CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Label-specific local metric learning improves nearest-neighbor multi-label classification by adding label- and region-specific private transformations to a shared one. The regions assigned to a training anchor, its support, select the anchor's transformation and also determine which private transformations its training constraints update. Existing methods partition the data, forcing instances with overlapping semantics into one region. Soft cluster memberships avoid this choice but send every anchor's supervision to every region; with exactly uniform memberships and positive private regularization, the private transformations of a label coincide at stationary points. We therefore restrict each local metric to the anchors that provide enough evidence for it. Granularity-Gated Fuzzy Metric Learning (GFML) sets this restriction with a gate derived from the principle of justifiable granularity. Each region receives a granule whose radius balances rival-adjusted membership coverage against specificity. An anchor inside a granule uses and trains only the private transformations whose granules contain it, so excluded transformations receive none of its supervision. On eight benchmarks with two KNN classifiers, GFML matches or improves on the strongest metric-learning and specialized baselines; ablations show that the gate improves on dense support.

1 INTRODUCTION

Multi-label classification (MLC) assigns several labels to each instance (Zhang & Zhou, 2014; Xu et al., 2016). Nearest-neighbor MLC classifiers need a similarity that adapts to each label, since instances can be close for one label and far for another, and to each region, since neighborhoods change across feature space. Global metric learning adapts to the labels but applies one geometry everywhere (Xing et al., 2002; Mao et al., 2023). Label-specific local metric learning (Mao et al., 2024b) handles both by combining, for each anchor, a shared transformation with a label-specific private transformation of its region.

We call the regions assigned to an anchor its *support*. The support plays two roles: it selects the transformation under which the anchor is compared with other instances, and, because each pairwise constraint is evaluated under its anchor's transformation, it decides which private transformations the anchor's constraints train. Partition-based local metric learners (Weinberger & Saul, 2009; Mao et al., 2024b) design the support for the first role and leave the second implicit. In multi-label data the two roles conflict, because overlapping semantics favor sharing components whereas specialized components need exclusive supervision.

Consider the City Park image in Figure 1, which shares attributes with Urban and Nature. A partition places it in one region, so its transformation jumps at the boundary and its constraints train a single private transformation although its neighborhood is mixed. Fuzzy C-Means (FCM) memberships (Bezdek et al., 1984) avoid this choice but are positive for every prototype, so City Park also trains the distant Ocean transformation, and every private transformation is trained by every anchor. With exactly uniform memberships and positive private regularization, each label's private transformations coincide at stationary points. FCM can approach this regime in high dimensions as its prototypes drift toward the data mean (Winkler et al., 2011); on four benchmarks, dense support

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

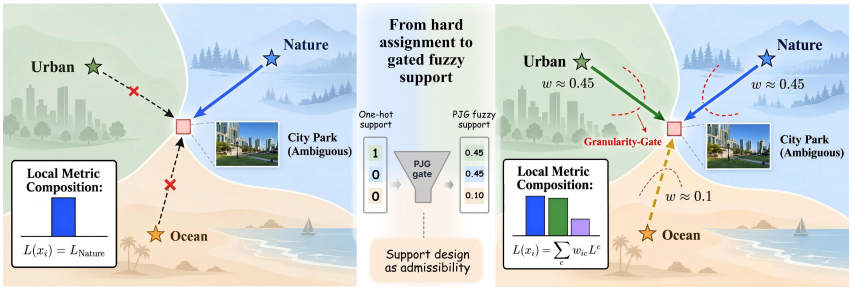


Figure 1: From hard assignment to gated fuzzy support. Left: City Park is assigned to Nature. Right: candidate fuzzy support before gating, with the subsequent granule gate overlaid. The displayed weights and mixture describe the fuzzy stage; the gate then restricts this support.

yields nearly identical private factors. Support design should therefore decide which anchors have enough evidence to use and train each local metric.

Granularity-Gated Fuzzy Metric Learning (GFML) makes this decision with a gate derived from the principle of justifiable granularity (PJG) (Pedrycz & Homenda, 2013): a granule should be supported by enough data (coverage) yet stay precise (specificity). For each FCM prototype, we pick among observed distances the radius maximizing coverage times specificity. Coverage credits each enclosed anchor with its membership margin over its strongest rival, so anchors that favor another prototype count against enlarging the granule. An anchor uses and trains only the private transformations whose granules contain it, one in a granule core and several where granules overlap; an anchor outside every granule keeps its FCM memberships instead of being forced into a region. Excluded transformations therefore receive no data gradient from a covered anchor. The gate overlaid in Figure 1 acts on the candidate fuzzy support: after gating, only the Urban and Nature private transformations receive City Park's anchor constraints.

PJG was introduced to build information granules that describe data; in GFML, granules instead decide which anchors train each local metric. Because the gate is fixed from the training features before metric learning, each anchor's transformation stays affine in the learned factors and every admissible region is explicit. Our contributions are:

- A support-selection formulation of label-specific local metric learning in which the same weights compose each anchor's transformation and route its constraints; hard partitions and dense fuzzy mixtures are special cases, and we show that dense mixtures with uniform memberships lose their locality.
- A justifiable-granularity gate with rival-adjusted coverage, to our knowledge the first use of justifiable granularity in metric learning. In the evaluated near-uniform regime, it preserves distinct private transformations where dense support exhibits near-collapse.
- An evaluation on eight benchmarks with two KNN classifiers and six metrics, with ablations that separate private capacity from the effect of the gate.

2 Related Work

Multi-Label Classification. Methods treat labels independently (Zhang & Zhou, 2014; Liu et al., 2022; Boutell et al., 2004), model pairwise dependencies (Yu & Zhang, 2022), or model dependencies across label subsets (Si et al., 2024). Other work studies surrogate-loss consistency (Mao et al., 2024a), ranking learnability (Raman et al., 2023), and selective label correlations (Chen et al., 2024). BR-KNN and ML-KNN (Boutell et al., 2004; Zhang & Zhou, 2007) predict from retrieved neighbors, making their performance sensitive to distance. GFML learns label-specific local geometry for these classifiers.

Multi-Label Metric Learning. Distance metric learning uses labeled supervision to construct task-specific similarities (Xing et al., 2002). Multi-label extensions include LM (Liu & Tsang, 2015), which aligns feature- and label-space similarity; LJE (Gouk et al., 2016), which approximates label Jaccard distance; COMMU (Sun & Zhang, 2021), which incorporates feature–label interactions; and

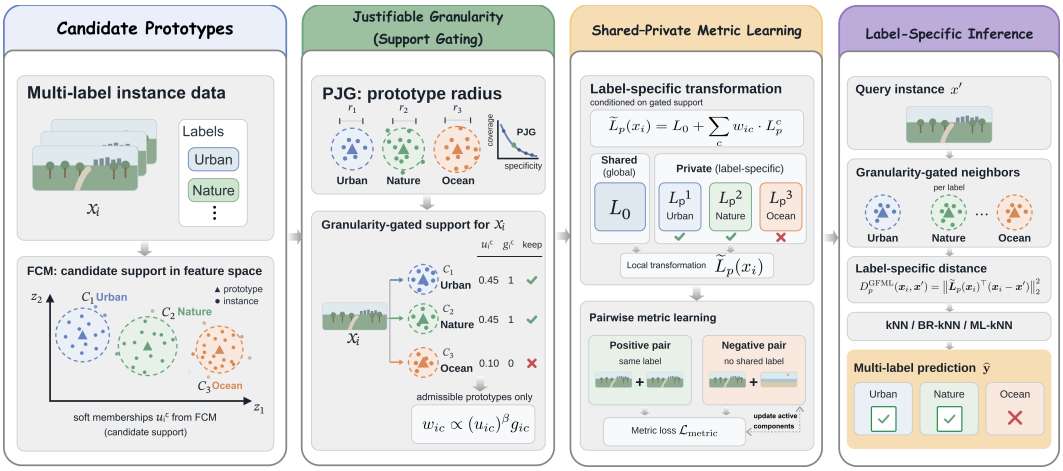


Figure 2: GFML workflow. FCM supplies candidate prototypes and memberships; justifiable granularity selects a radius for each prototype and gates the support of each training anchor. The resulting weights compose its shared–private transformation and route pairwise supervision to admitted private factors. At inference, label-specific distances from training anchors to a query determine the neighbors used by the KNN classifier.

LIMIC (Mao et al., 2023), which regularizes label-specific metrics through co-occurrence. These approaches use metrics that are constant across the instance space. LSMM (Mao et al., 2024b) introduces local adaptation through a shared transformation and label-specific private components defined by a partition. GFML keeps this shared–private representation but replaces the partition with a support that also controls which anchors supervise each private component.

Local Supports and Granules. Multiple-metric LMNN (Weinberger & Saul, 2009) and LSMM train one metric per region of a partition. FCM (Bezdek et al., 1984) gives graded prototype memberships, and multi-label methods also use prototypes for angular margins (Li et al., 2024) or label alignment (Lv et al., 2025). PJG (Pedrycz & Homenda, 2013) builds information granules that balance coverage and specificity. GFML uses such granules, fixed before metric learning, to select the anchors that supervise each label-specific local metric.

3 Preliminaries

3.1 Notation and Problem Definition

Let $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i)\}_{i=1}^n$ be a multi-label dataset with features $\boldsymbol{x}_i \in \mathbb{R}^d$ and relevant labels $Y_i \subseteq \mathcal{Y} = \{l_1, \ldots, l_q\}$. Write $\boldsymbol{y}_i \in \{0,1\}^q$ for its indicator vector, with $y_i^p = 1$ exactly when $l_p \in Y_i$. We seek a classifier $h : \mathbb{R}^d \to 2^{\mathcal{Y}}$ for unseen instances.

A positive semidefinite (PSD) matrix $\boldsymbol{M} \in \mathbb{S}_+^d$ defines a Mahalanobis distance. Factoring $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^\top$ with $\boldsymbol{L} \in \mathbb{R}^{d \times d'}$ ($d' \leq d$) gives

$$\mathrm{Dis}_{\boldsymbol{L}}^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j) = \|\boldsymbol{L}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_2^2. \tag{1}$$

This parameterization preserves positive semidefiniteness without a projection step.

3.2 Label-Specific Pairwise Constraints

For label l_p, define the positive set $\mathcal{P}_p = \{\boldsymbol{x}_i \mid y_i^p = 1\}$ and negative set $\mathcal{N}_p = \{\boldsymbol{x}_i \mid y_i^p = 0\}$. The constraint set $\mathcal{T}_p = \{(i, j, \theta_{ij}^p)\}$ contains ordered pairs, with i the anchor, j the neighbor, and

$$\theta_{ij}^p = \begin{cases} +1, & \text{if } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{P}_p \times \mathcal{P}_p \cup \mathcal{N}_p \times \mathcal{N}_p, \\ -1, & \text{otherwise.} \end{cases} \tag{2}$$

Each constraint uses the anchor's local transformation. We sample similar targets and dissimilar imposters from each anchor's nearest neighbors (Weinberger et al., 2005).

4 THE GFML FRAMEWORK

GFML first constructs support from the training features, then learns shared and label-specific private transformations under that support. Figure 2 summarizes the route from candidate prototypes and granule radii to local metric learning and label-specific inference.

4.1 LABEL-SPECIFIC LOCAL METRIC AGGREGATION

Let $\boldsymbol{L}_0 \in \mathbb{R}^{d\times d'}$ be shared across labels and regions, and let $\boldsymbol{L}_p^c \in \mathbb{R}^{d\times d'}$ be the private transformation for label p and region c. A hard region assignment $c(\boldsymbol{x}_i)$ gives the shared–private form

$$\boldsymbol{L}_p^{\mathrm{hard}}(\boldsymbol{x}_i) = \boldsymbol{L}_0 + \boldsymbol{L}_p^{c(\boldsymbol{x}_i)}. \tag{3}$$

GFML replaces the single private component with a mixture. Given nonnegative support weights $\boldsymbol{W} = [w_{ic}] \in \mathbb{R}^{n\times C}$ with $\sum_c w_{ic} = 1$, define

$$\widetilde{\boldsymbol{L}}_p(\boldsymbol{x}_i) = \boldsymbol{L}_0 + \sum_{c=1}^{C} w_{ic}\boldsymbol{L}_p^c. \tag{4}$$

One-hot weights recover Eq. 3; fuzzy weights allow several components to contribute. The support is shared across labels, while the private transformations are label-specific. Aggregation takes place in the transformation factors, inducing the PSD matrix $\widetilde{\boldsymbol{M}}_p(\boldsymbol{x}_i) = \widetilde{\boldsymbol{L}}_p(\boldsymbol{x}_i)\,\widetilde{\boldsymbol{L}}_p(\boldsymbol{x}_i)^\top$.

The squared distance between anchor $\boldsymbol{x}_i$ and neighbor $\boldsymbol{x}_j$ under label p is

$$\mathrm{Dis}_p^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = \big\|\widetilde{\boldsymbol{L}}_p(\boldsymbol{x}_i)^\top(\boldsymbol{x}_i - \boldsymbol{x}_j)\big\|_2^2. \tag{5}$$

4.2 GRANULARITY-GATED FUZZY SUPPORT

FCM (Bezdek et al., 1984) produces C prototypes $\{\boldsymbol{v}_c\}_{c=1}^C \subset \mathbb{R}^d$ and candidate memberships $\boldsymbol{U} = [u_{ic}] \in [0,1]^{n\times C}$. PJG (Pedrycz & Homenda, 2013) then assigns a radius ρ_c^* to each prototype by balancing coverage and specificity. Define the prototype distance

$$d_c(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{v}_c\|_2. \tag{6}$$

For the granule $\mathcal{A}_c(\rho) = \{\boldsymbol{x} : d_c(\boldsymbol{x}) \le \rho\}$, coverage measures the membership advantage over each anchor's strongest rival:

$$\mathrm{cov}_c(\rho) = \max\left(0, \frac{1}{n}\sum_{\boldsymbol{x}_i\in\mathcal{A}_c(\rho)}\big(u_{ic} - u_{i,r_c(i)}\big)\right), \tag{7}$$

where $r_c(i) = \arg\max_{k\neq c} u_{ik}$. Anchors that favor prototype c contribute positively, while those that favor a rival reduce the accumulated support. Specificity favors a compact region:

$$\mathrm{sp}_c(\rho) = 1 - \frac{\rho}{D_c}, \quad \text{where } D_c = \max_i d_c(\boldsymbol{x}_i). \tag{8}$$

We select the radius from the observed distances,

$$\mathcal{S}_c = \big\{d_c(\boldsymbol{x}_i) \;\big|\; i = 1,\dots,n\big\} \cup \{0\}, \tag{9}$$

by maximizing

$$\rho_c^* = \arg\max_{\rho\in\mathcal{S}_c}\Big(\mathrm{cov}_c(\rho)\cdot\mathrm{sp}_c(\rho)\Big). \tag{10}$$

The criterion favors regions that retain membership support without extending far from their prototype. It determines the radius from the training geometry, with no additional radius threshold to tune. The finite search suffices because coverage changes only at observed distances and specificity decreases between them. The supplementary material details the FCM step and radius search.

The radius defines a binary admissibility gate,

$$g_{ic} = \mathbb{I}\big[\|\boldsymbol{x}_i - \boldsymbol{v}_c\|_2 \le \rho_c^*\big], \tag{11}$$

with $\mathbb{I}[\cdot]$ the indicator. The raw gated weights are

$$\psi_{ic} = (u_{ic})^{\beta} \cdot g_{ic}, \tag{12}$$

where $\beta \geq 1$ controls concentration within the active set. Normalization gives

$$w_{ic} = \begin{cases} \dfrac{\psi_{ic}}{\sum_{k=1}^{C} \psi_{ik}}, & \text{if } \sum_{k=1}^{C} \psi_{ik} > 0, \\ u_{ic}, & \text{otherwise.} \end{cases} \tag{13}$$

For an anchor covered by at least one granule, only admitted prototypes have positive weights, and overlapping granules permit a fuzzy mixture. If no granule covers the anchor, Eq. 13 retains its dense FCM memberships. Thus every anchor has a defined transformation.

The radius and β have distinct roles. The gate selects admissible components, while β controls their relative influence. For two active components with positive memberships,

$$\frac{w_{ic}}{w_{ik}} = \left(\frac{u_{ic}}{u_{ik}}\right)^{\beta}. \tag{14}$$

At $\beta = 1$, GFML preserves the original membership ratios within the active set. Larger values concentrate the mixture toward its highest-membership component without changing which components are admitted.

4.3 Objective Function and Optimization

For each ordered constraint $(i, j, \theta_{ij}^{p}) \in \mathcal{T}_p$ with anchor i, define the margin variable

$$z_{ij}^{p} = \theta_{ij}^{p} \left(\gamma - \alpha \, \mathrm{Dis}_p^2(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), \tag{15}$$

where $\gamma > 1$ and $\alpha > 0$ set the distance margins. We use the continuously differentiable smooth hinge loss

$$\ell(z) = \begin{cases} 0, & \text{if } z > 1, \\ \frac{1}{2}(z-1)^2, & \text{if } 0 \leq z \leq 1, \\ \frac{1}{2} - z, & \text{if } z < 0. \end{cases} \tag{16}$$

The full objective is

$$\min_{\boldsymbol{L}_0, \{\boldsymbol{L}_p^c\}} \mathcal{J} = \frac{1}{q} \sum_{p=1}^{q} \frac{1}{|\mathcal{T}_p|} \sum_{(i,j,\theta_{ij}^{p}) \in \mathcal{T}_p} \ell(z_{ij}^{p}) + \frac{\lambda_1}{2Cq} \sum_{p=1}^{q} \sum_{c=1}^{C} \|\boldsymbol{L}_p^c\|_F^2 + \frac{\lambda_2}{2} \|\boldsymbol{L}_0\|_F^2, \tag{17}$$

where λ_1 and λ_2 regularize the private and shared transformations, respectively. We jointly optimize these transformations using limited-memory BFGS (L-BFGS) (Liu & Nocedal, 1989). The supplementary material provides the gradient derivation and training algorithm.

We hold $\boldsymbol{W}$ fixed during metric learning. The effective transformation is then affine in the learned factors, separating support construction from supervised metric fitting. This two-stage design uses the input-space geometry as a structural prior; the support itself does not adapt during optimization.

4.4 How Support Governs Local Learning

The support weights act on both the transformation and its supervision. Let $\mathcal{T}_p(i)$ collect the constraints anchored at i, and write $\boldsymbol{G}_{ij}^{p} = \partial \ell(z_{ij}^{p}) / \partial \widetilde{\boldsymbol{L}}_p(\boldsymbol{x}_i)$. By the chain rule, the gradient of a private factor is

$$\frac{\partial \mathcal{J}}{\partial \boldsymbol{L}_p^c} = \frac{1}{q|\mathcal{T}_p|} \sum_{i=1}^{n} w_{ic} \sum_{(i,j,\theta_{ij}^{p}) \in \mathcal{T}_p(i)} \boldsymbol{G}_{ij}^{p} + \frac{\lambda_1}{Cq} \boldsymbol{L}_p^c. \tag{18}$$

Thus an anchor with $w_{ic} = 0$ contributes no data gradient to $\boldsymbol{L}_p^c$. One-hot support assigns its constraints to a single private factor. Dense fuzzy support distributes them across all components

Table 1: Statistics of the benchmark datasets.

	emotions	birds	medical	enron	image	scene	arts	education
Instances	593	645	978	1702	2000	2407	5000	5000
Features	72	258	1449	1001	294	294	462	550
Labels	6	19	45	53	5	6	26	33
Label cardinality	1.869	1.014	1.245	3.378	1.236	1.074	1.636	1.461
Domain	Music	Audio	Text	Text	Image	Image	Text	Text

with positive memberships. Gated support restricts this distribution to the admitted components for covered anchors, while permitting sharing where granules overlap. The shared factor $\boldsymbol{L}_0$ continues to receive supervision from all labels and anchors.

Dense support can erase locality. If $w_{ic} = 1/C$ for all i and c, every anchor uses $\boldsymbol{L}_0 + \frac{1}{C}\sum_c \boldsymbol{L}_p^c$, so all private factors of label p share the data term of Eq. 18; for $\lambda_1 > 0$, setting Eq. 18 to zero gives $\boldsymbol{L}_p^1 = \cdots = \boldsymbol{L}_p^C$ at any stationary point, leaving one metric per label. FCM can approach this limit in high dimensions (Winkler et al., 2011). Distinct admission sets can break the equality of the private data-gradient terms. In the evaluated near-uniform regime, dense support exhibits near-zero dispersion in every fold on *scene*, *image*, *medical*, and *enron*, while GFML retains distinct private factors. This observation does not imply recovery at exact uniformity. Section 5.3 separates the gate's effect from private capacity.

4.5 Classification Inference

For a query $\boldsymbol{x}'$, we score each training anchor $\boldsymbol{x}_i$ using its learned transformation:

$$D_p^{\mathrm{GFML}}(\boldsymbol{x}_i, \boldsymbol{x}') = \left\| \widetilde{\boldsymbol{L}}_p(\boldsymbol{x}_i)^\top (\boldsymbol{x}_i - \boldsymbol{x}') \right\|_2^2. \tag{19}$$

For each label, we retrieve the k training anchors with the smallest scores and apply BR-KNN or ML-KNN. The dissimilarity is generally asymmetric because the transformation belongs to the training anchor. This matches the directed training constraints and requires no support estimation for a new query.

5 Experiments

5.1 Experimental Setup

Table 1 summarizes the eight benchmarks, which span music, audio, text, and image data. We apply PCA at 95% variance to all datasets except *emotions* and *birds*. For the benchmark comparisons and component ablations, we hold out 10% of each dataset as a validation set for grid-search hyperparameter selection, then run 10-fold cross-validation on the remaining 90% with those hyperparameters fixed, reporting mean±std across folds. The same protocol is used for every baseline. We report six standard multi-label metrics (Zhang & Zhou, 2014): *Hamming Loss*, *Ranking Loss* and *Coverage* (↓), and *Average Precision*, *Macro-F1*, *Macro-AUC* (↑).

GFML uses the common support settings $C = 3$, $m = 2$, and $\beta = 1$ across datasets. We select λ_1 and λ_2 by validation. The supplementary material lists the hyperparameter grids, baseline configurations, and hardware.

5.2 Comparative Studies

We evaluate the learned geometry through BR-KNN and ML-KNN. The metric-learning baselines are LM (Liu & Tsang, 2015), LJE (Gouk et al., 2016), COMMU (Sun & Zhang, 2021), LIMIC (Mao et al., 2023), and the hard local method LSMM-CL (Mao et al., 2024b). We also compare with four specialized multi-label classifiers: LIFT (Zhang & Wu, 2015), RELIAB (Zhang et al., 2021), WRAP (Yu & Zhang, 2022), and HOMI (Si et al., 2024).

Results against metric learning baselines. Tables 2 and 3 report *Hamming Loss*, *Ranking Loss*, and *Macro-F1*; the remaining metrics are reported in the supplementary material. Across the six

Table 2: Hamming Loss and Ranking Loss comparison with metric-learning baselines (mean±std across folds). Bold and underlined entries mark the best and second-best reported means within each KNN classifier, including ties.

Hamming Loss ↓

Compared Algorithms	emotions	birds	medical	enron	image	scene	arts	education
BR-KNN	0.263±0.023	0.056±0.007	0.016±0.002	0.055±0.002	0.167±0.016	0.088±0.008	0.073±0.002	**0.038±0.001**
BR-KNN-LM	0.270±0.019	0.065±0.009	**0.011±0.002**	0.048±0.003	0.180±0.016	0.088±0.012	0.055±0.002	**0.038±0.001**
BR-KNN-LJE	0.219±0.022	0.055±0.007	0.022±0.003	0.059±0.003	0.184±0.018	0.110±0.011	0.061±0.001	0.043±0.002
BR-KNN-COMMU	0.263±0.023	0.056±0.007	0.016±0.002	0.055±0.002	0.167±0.016	0.088±0.008	0.072±0.002	**0.038±0.001**
BR-KNN-LIMIC	0.212±0.008	0.053±0.006	0.014±0.002	0.049±0.003	0.165±0.016	0.083±0.008	0.058±0.002	0.039±0.001
BR-KNN-LSMM-CL	0.202±0.017	**0.050±0.007**	0.012±0.002	0.044±0.003	**0.157±0.017**	0.079±0.009	**0.052±0.002**	**0.038±0.001**
BR-KNN-GFML (Ours)	**0.197±0.028**	**0.050±0.006**	**0.011±0.003**	**0.043±0.002**	**0.157±0.013**	**0.076±0.008**	0.053±0.002	**0.038±0.001**
ML-KNN	0.262±0.022	0.054±0.006	0.016±0.002	0.052±0.003	0.171±0.013	0.084±0.009	0.060±0.001	**0.038±0.001**
ML-KNN-LM	0.254±0.017	0.054±0.007	0.013±0.002	0.048±0.002	0.174±0.015	0.086±0.010	0.054±0.001	**0.038±0.001**
ML-KNN-LJE	0.227±0.022	0.054±0.007	0.023±0.003	0.058±0.002	0.184±0.017	0.109±0.009	0.060±0.001	0.042±0.002
ML-KNN-COMMU	0.262±0.022	0.054±0.006	0.015±0.002	0.052±0.003	0.171±0.013	0.084±0.009	0.059±0.001	**0.038±0.001**
ML-KNN-LIMIC	0.215±0.009	0.053±0.006	0.013±0.002	0.050±0.003	0.164±0.017	0.082±0.007	0.057±0.002	**0.038±0.002**
ML-KNN-LSMM-CL	0.209±0.018	**0.051±0.006**	**0.012±0.002**	**0.045±0.003**	**0.155±0.013**	0.078±0.007	**0.052±0.001**	**0.038±0.001**
ML-KNN-GFML (Ours)	**0.201±0.028**	**0.051±0.005**	**0.012±0.002**	0.048±0.003	0.158±0.013	**0.076±0.006**	0.053±0.002	**0.038±0.001**

Ranking Loss ↓

Compared Algorithms	emotions	birds	medical	enron	image	scene	arts	education
BR-KNN	0.272±0.048	0.516±0.064	0.081±0.028	0.228±0.023	0.180±0.020	0.095±0.014	0.354±0.029	0.244±0.013
BR-KNN-LM	0.256±0.030	0.527±0.066	0.105±0.033	0.201±0.011	0.191±0.020	0.097±0.016	0.257±0.015	0.222±0.014
BR-KNN-LJE	0.202±0.035	0.497±0.067	0.114±0.037	0.244±0.027	0.205±0.020	0.124±0.019	0.265±0.015	0.234±0.012
BR-KNN-COMMU	0.272±0.048	0.516±0.064	0.088±0.031	0.230±0.023	0.180±0.020	0.095±0.014	0.357±0.033	0.244±0.013
BR-KNN-LIMIC	0.182±0.028	0.513±0.066	0.077±0.023	0.245±0.020	0.171±0.024	0.102±0.011	0.348±0.021	0.220±0.015
BR-KNN-LSMM-CL	0.178±0.034	0.487±0.058	**0.071±0.023**	**0.192±0.022**	**0.150±0.025**	0.090±0.013	0.258±0.027	**0.205±0.012**
BR-KNN-GFML (Ours)	**0.172±0.035**	**0.395±0.056**	0.084±0.031	**0.192±0.017**	0.168±0.028	**0.087±0.011**	**0.254±0.015**	0.221±0.019
ML-KNN	0.258±0.038	0.295±0.035	0.032±0.009	0.091±0.010	0.173±0.018	0.075±0.011	0.148±0.008	0.097±0.004
ML-KNN-LM	0.240±0.026	0.298±0.047	0.034±0.015	0.090±0.008	0.177±0.018	0.078±0.012	0.129±0.007	0.078±0.005
ML-KNN-LJE	0.201±0.030	0.299±0.047	0.053±0.017	0.111±0.011	0.201±0.022	0.110±0.017	0.145±0.009	0.089±0.003
ML-KNN-COMMU	0.258±0.038	0.295±0.035	0.038±0.013	0.091±0.010	0.173±0.018	0.075±0.011	0.147±0.008	0.097±0.004
ML-KNN-LIMIC	0.184±0.020	0.263±0.039	**0.031±0.012**	0.088±0.010	0.163±0.022	0.072±0.007	0.137±0.006	0.079±0.004
ML-KNN-LSMM-CL	0.170±0.036	0.260±0.038	0.038±0.013	**0.073±0.009**	**0.146±0.020**	0.069±0.005	0.120±0.005	**0.074±0.004**
ML-KNN-GFML (Ours)	**0.166±0.032**	**0.259±0.036**	0.035±0.013	0.085±0.007	0.154±0.020	**0.062±0.008**	**0.119±0.005**	0.078±0.005

Table 3: Macro-F1 comparison with metric-learning baselines (mean±std across folds). Bold and underline as in Table 2.

Macro-F1 ↑

Compared Algorithms	emotions	birds	medical	enron	image	scene	arts	education
BR-KNN	0.463±0.045	0.038±0.018	0.169±0.018	0.092±0.013	0.599±0.033	0.731±0.016	0.126±0.017	0.137±0.015
BR-KNN-LM	0.490±0.033	0.060±0.035	0.320±0.051	0.166±0.020	0.597±0.034	0.744±0.035	0.191±0.009	0.162±0.017
BR-KNN-LJE	0.583±0.047	0.034±0.016	0.273±0.027	0.061±0.012	0.554±0.040	0.664±0.028	0.148±0.014	0.122±0.015
BR-KNN-COMMU	0.463±0.045	0.047±0.025	0.297±0.040	0.109±0.015	0.599±0.033	0.731±0.016	0.136±0.017	0.176±0.023
BR-KNN-LIMIC	0.609±0.016	0.093±0.028	0.384±0.031	0.142±0.020	0.622±0.034	0.760±0.024	0.185±0.019	0.157±0.016
BR-KNN-LSMM-CL	0.620±0.033	0.101±0.029	0.526±0.048	**0.178±0.023**	**0.638±0.039**	0.757±0.023	**0.213±0.019**	0.179±0.021
BR-KNN-GFML (Ours)	**0.631±0.053**	**0.110±0.026**	**0.593±0.032**	0.177±0.014	0.628±0.031	**0.773±0.021**	0.178±0.019	**0.197±0.030**
ML-KNN	0.361±0.050	0.010±0.007	0.210±0.027	0.086±0.013	0.581±0.028	0.748±0.022	0.098±0.010	0.126±0.023
ML-KNN-LM	0.438±0.036	0.016±0.011	0.294±0.041	0.146±0.010	0.573±0.031	0.746±0.031	0.187±0.022	0.161±0.028
ML-KNN-LJE	0.530±0.052	0.009±0.011	0.245±0.044	0.045±0.012	0.530±0.037	0.651±0.023	0.125±0.012	0.101±0.018
ML-KNN-COMMU	0.361±0.050	0.013±0.008	0.364±0.061	0.102±0.009	0.581±0.028	0.748±0.022	0.108±0.013	0.161±0.035
ML-KNN-LIMIC	0.587±0.028	0.071±0.030	0.384±0.040	0.134±0.020	0.635±0.031	0.763±0.019	0.197±0.018	0.182±0.032
ML-KNN-LSMM-CL	0.597±0.027	0.097±0.037	**0.551±0.049**	**0.165±0.019**	**0.649±0.035**	0.761±0.027	**0.226±0.024**	0.207±0.027
ML-KNN-GFML (Ours)	**0.609±0.047**	**0.098±0.023**	0.537±0.063	0.159±0.018	0.641±0.026	**0.780±0.013**	0.198±0.016	**0.216±0.036**

metrics, eight datasets, and two classifiers, GFML has the best mean in 53 of the 96 comparisons (ties included). With BR-KNN, it matches or improves on LSMM-CL in Hamming Loss on seven datasets. Gains also appear in ranking. On *birds*, BR-KNN Ranking Loss decreases from 0.487 with LSMM-CL to 0.395 with GFML, and on *scene*, ML-KNN Ranking Loss decreases from 0.069 to 0.062.

Results against non-metric learning baselines. Table 4 compares GFML-equipped KNN with specialized classifiers. The two GFML variants take the best Macro-F1 mean on seven of eight

Table 4: Comparison with specialized multi-label classifiers (mean±std across folds). Bold and underline as in Table 2.

Compared Algorithms	emotions	birds	medical	enron	image	scene	arts	education
Hamming Loss ↓								
LIFT	0.267±0.015	0.074±0.026	0.012±0.002	0.046±0.002	0.159±0.014	0.079±0.007	**0.053±0.001**	**0.037±0.001**
RELIAB	0.260±0.031	0.074±0.007	0.016±0.001	0.059±0.003	0.178±0.015	0.110±0.015	0.055±0.003	**0.037±0.002**
WRAP	0.242±0.019	0.053±0.008	**0.010±0.001**	0.047±0.002	0.186±0.009	0.114±0.007	0.054±0.002	**0.037±0.001**
HOMI	0.238±0.018	0.073±0.025	0.013±0.002	0.045±0.002	0.162±0.017	0.085±0.006	0.057±0.002	0.038±0.001
BR-KNN-GFML (Ours)	**0.197±0.028**	**0.050±0.006**	0.011±0.003	**0.043±0.002**	**0.157±0.013**	**0.076±0.008**	**0.053±0.002**	0.038±0.001
ML-KNN-GFML (Ours)	0.201±0.028	0.051±0.005	0.012±0.002	0.048±0.003	0.158±0.013	**0.076±0.006**	**0.053±0.002**	0.038±0.001
Ranking Loss ↓								
LIFT	0.254±0.046	0.323±0.050	0.028±0.011	**0.079±0.006**	0.156±0.020	0.074±0.009	0.124±0.006	0.084±0.003
RELIAB	0.256±0.040	0.427±0.036	**0.014±0.009**	0.085±0.009	0.181±0.025	0.091±0.009	0.145±0.006	0.097±0.008
WRAP	0.229±0.029	0.322±0.030	0.017±0.008	0.090±0.009	0.173±0.023	0.085±0.008	0.128±0.009	0.086±0.006
HOMI	0.237±0.032	0.329±0.046	0.041±0.012	0.098±0.013	0.196±0.021	0.084±0.010	0.186±0.010	0.132±0.004
BR-KNN-GFML (Ours)	0.172±0.035	0.395±0.056	0.084±0.031	0.192±0.017	0.168±0.028	0.087±0.011	0.254±0.015	0.221±0.019
ML-KNN-GFML (Ours)	**0.166±0.032**	**0.259±0.036**	0.035±0.013	0.085±0.007	**0.154±0.020**	**0.062±0.008**	**0.119±0.005**	**0.078±0.005**
Macro-F1 ↑								
LIFT	0.432±0.040	0.024±0.018	0.310±0.061	0.150±0.015	0.606±0.036	0.759±0.016	0.123±0.009	0.156±0.011
RELIAB	0.538±0.047	0.098±0.012	0.473±0.038	0.132±0.008	0.555±0.016	0.687±0.015	0.151±0.003	0.176±0.004
WRAP	0.318±0.036	0.093±0.012	0.529±0.022	0.148±0.006	0.501±0.036	0.773±0.029	0.192±0.012	0.194±0.005
HOMI	0.488±0.028	0.068±0.023	0.511±0.061	0.126±0.006	**0.650±0.034**	0.759±0.025	0.186±0.020	0.183±0.016
BR-KNN-GFML (Ours)	**0.631±0.053**	**0.110±0.026**	**0.593±0.032**	**0.177±0.014**	0.628±0.031	0.773±0.021	0.178±0.019	0.197±0.030
ML-KNN-GFML (Ours)	0.609±0.047	0.098±0.023	0.537±0.063	0.159±0.018	0.641±0.026	**0.780±0.013**	**0.198±0.016**	**0.216±0.036**

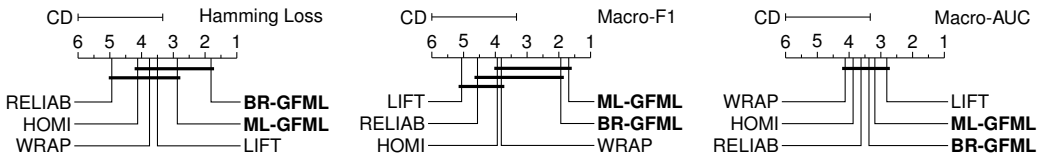


Figure 3: Friedman average ranks over eight datasets for the GFML variants and the specialized classifiers (1 is best). Bars join methods that the Nemenyi test does not separate ($\alpha = 0.05$, $\mathrm{CD} = 2.67$). BR-GFML and ML-GFML denote BR-KNN-GFML and ML-KNN-GFML.

datasets, and ML-KNN-GFML attains the lowest Ranking Loss on six. The two KNN rules differ in emphasis. BR-KNN-GFML is strong on Hamming Loss, whereas ML-KNN-GFML gives lower Ranking Loss than BR-KNN-GFML on every dataset. Learning the neighborhood geometry therefore provides a competitive alternative to more specialized classifier designs.

Statistical significance analysis. Figure 3 compares average ranks against the specialized classifiers with the Friedman–Nemenyi test. BR-KNN-GFML has the best Hamming Loss rank and ML-KNN-GFML the best Macro-F1 rank, and both variants rank significantly above LIFT on Macro-F1. Results for the other metrics are provided in the supplementary material.

5.3 Contribution of Local Aggregation and Gating

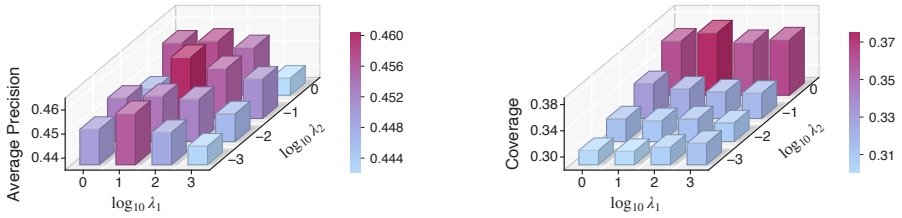
Table 5 separates local capacity from support selection. The shared-only variant learns $\boldsymbol{L}_0$ without private components; the dense fuzzy variant retains the private mixture without granule gating. On all three datasets, adding private components and then gating their support progressively reduces mean Hamming Loss. Relative to dense fuzzy support, GFML reduces it by 0.008, 0.002, and 0.006 on *emotions*, *medical*, and *scene*, respectively. On *scene*, GFML also improves on dense support at each tested neighborhood size, $k \in \{3, 5, 10, 15, 20\}$. The supplementary material documents near-collapse of the dense variant's private factors on datasets where FCM memberships approach uniformity.

5.4 Parameter Sensitivity

C sets the number of prototypes, β sharpens weights among admissible components, and m controls the FCM memberships from which support is constructed; we vary them in BR-KNN sweeps with

Table 5: Ablation study of GFML. Hamming Loss using BR-KNN (mean±std, ↓).

Variant	emotions	medical	scene
GFML	**0.197±0.028**	**0.011±0.003**	**0.076±0.008**
Dense fuzzy (no gate)	0.205±0.019	0.013±0.002	0.082±0.009
Shared only	0.225±0.021	0.014±0.002	0.085±0.010



(a) birds, Average Precision (↑)

(b) emotions, Coverage (↓)

Figure 4: Regularization sensitivity with BR-KNN over $\log_{10} \lambda_1$ (horizontal) and $\log_{10} \lambda_2$ (receding).

the remaining settings and folds fixed. Table 6 shows that the results are insensitive to C and that increasing β gives no consistent advantage over $\beta = 1$. The fuzzifier m shows stronger dataset dependence; the sweep and its diagnostic protocol are reported in the supplementary material.

Table 6: Support-parameter sensitivity. Hamming Loss with BR-KNN (mean±std across ten folds, ↓). The C sweep fixes $m = 2$, $\beta = 1$; the β sweep fixes $C = 3$, $m = 2$.

Dataset	$C = 2$	$C = 3$	$C = 4$	$C = 5$
emotions	0.1998 ± 0.0277	0.2024 ± 0.0254	0.2007 ± 0.0287	0.2002 ± 0.0275
birds	0.0502 ± 0.0054	0.0514 ± 0.0065	0.0507 ± 0.0066	0.0515 ± 0.0072
medical	0.0094 ± 0.0023	0.0093 ± 0.0023	0.0091 ± 0.0023	0.0093 ± 0.0022
scene	0.0764 ± 0.0080	0.0778 ± 0.0074	0.0778 ± 0.0087	0.0801 ± 0.0069
Dataset	$\beta = 1$	$\beta = 2$	$\beta = 4$	$\beta = 8$
emotions	0.2024 ± 0.0254	0.2010 ± 0.0256	0.2024 ± 0.0273	0.2007 ± 0.0261

Regularization. Figure 4 varies the penalties on the private (λ_1) and shared (λ_2) transformations. On *birds*, Average Precision peaks near $\lambda_1 = 10$, $\lambda_2 = 10^{-1}$ and generally decreases at $\lambda_1 = 10^3$; on *emotions*, Coverage is lower for smaller λ_2 across all λ_1. The trends differ by dataset and metric, so we select both penalties by validation.

6 Conclusion

We introduced GFML to decide which private transformations each training example uses and trains. Its justifiable granules preserve fuzzy sharing across overlapping regions, while ensuring that excluded transformations receive no data gradient from a covered anchor. On eight multi-label benchmarks, the learned geometry makes standard KNN classifiers competitive with metric-learning and specialized methods. Our ablations show gains from both private transformations and gating. In the evaluated near-uniform regime, gating also keeps private factors distinct where dense fuzzy support nearly collapses.

AI use statement

Generative AI tools were used for language polishing (ChatGPT) and coding assistance (Claude Code). The authors remain responsible for the final manuscript and code, including AI-assisted material.

Ethics statement

Our experiments use existing multi-label benchmark datasets, including the *medical* and *enron* datasets, which originate from clinical text and email collections. We report aggregate evaluation results. The sources and available license information are listed in the supplementary material.

Reproducibility statement

The main text describes the model and optimization procedure in Sections 4 and 5. The supplementary material gives the gradient derivation, gating definitions, dataset sources, training algorithm, full experimental tables, and statistical comparisons. Source code, configurations, and scripts reproducing all tables will be released upon acceptance.

References

James C. Bezdek, Robert Ehrlich, and William Full. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203, 1984.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

Zhao-Min Chen, Xin Jin, Yisu Ge, and Sixian Chan. In pursuit of causal label correlations for multi-label image recognition. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Henry Gouk, Bernhard Pfahringer, and Michael Cree. Learning distance metrics for multi-label classification. In *Asian Conference on Machine Learning*, pp. 318–333. PMLR, 2016.

Ximing Li, Silong Liang, Changchun Li, Pengfei Wang, and Fangming Gu. Semi-supervised multi-label learning with balanced binary angular margin loss. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

Weiwei Liu and Ivor Tsang. Large margin metric learning for multi-label prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7955–7974, 2022.

Weijun Lv, Yu Chen, Xiaozhao Fang, Xuhuan Zhu, Jie Wen, Guoxu Zhou, and Sixian Chan. Confidence-aware with prototype alignment for partial multi-label learning. In *Advances in Neural Information Processing Systems*, 2025. NeurIPS 2025 poster.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-label learning with stronger consistency guarantees. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Jun-Xiang Mao, Jun-Yi Hang, and Min-Ling Zhang. Learning label-specific multiple local metrics for multi-label classification. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 4742–4750, 2024b.

Junxiang Mao, Wei Wang, and Min-Ling Zhang. Label specific multi-semantics metric learning for multi-label classification: Global consideration helps. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4055–4063, 2023.

Witold Pedrycz and Wladyslaw Homenda. Building the fundamentals of granular computing: A principle of justifiable granularity. *Applied Soft Computing*, 13(10):4209–4218, 2013.

Vinod Raman, Unique Subedi, and Ambuj Tewari. On the learnability of multilabel ranking. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chongjie Si, Yuheng Jia, Ran Wang, Min-Ling Zhang, Yanghe Feng, and Chongxiao Qu. Multi-label classification with high-rank and high-order label correlations. *IEEE Transactions on Knowledge and Data Engineering*, 36(8):4076–4088, 2024.

Yan-Ping Sun and Min-Ling Zhang. Compositional metric learning for multi-label classification. *Frontiers of Computer Science*, 15(5):155320, 2021.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

Kilian Q. Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Roland Winkler, Frank Klawonn, and Rudolf Kruse. Fuzzy C-means in high dimensional spaces. *International Journal of Fuzzy System Applications*, 1(1):1–16, 2011.

Eric Xing, Michael Jordan, Stuart J. Russell, and Andrew Ng. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

Suping Xu, Xibei Yang, Hualong Yu, Dong-Jun Yu, Jingyu Yang, and Eric C. C. Tsang. Multi-label learning with label-specific feature reduction. *Knowledge-Based Systems*, 104:52–61, 2016.

Ze-Bang Yu and Min-Ling Zhang. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 5199–5210, 2022.

Min-Ling Zhang and Lei Wu. LIFT: Multi-label learning with label-specific features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):107–120, 2015.

Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

Min-Ling Zhang, Qian-Wen Zhang, Jun-Peng Fang, Yu-Kun Li, and Xin Geng. Leveraging implicit relative labeling-importance information for effective multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):2057–2070, 2021.