DENO: SIGNED DETERMINANTAL EVIDENCE POOLING FOR CORRELATED MULTIPLE-INSTANCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple-instance learning (MIL) often involves strongly correlated or redundant instances, yet standard aggregation does not explicitly separate redundant evidence from prevalence information. We introduce DENO, a signed determinantal evidence pooling operator that combines signed instance evidence with a degree-normalized similarity kernel and contrasts log-determinants under opposite evidence tilts. A separate mean-evidence branch retains frequency information. We show that, for exact replicas, the full instance-level determinant reduces exactly to an effective latent-source determinant: uniform replication leaves the determinantal term unchanged, while selective replication can alter the effective source geometry. For near-redundant groups, we bound the approximation error using residual spectral structure and within-group evidence variation. We further develop a boundary-aware ordinal extension and evaluate DENO across five heterogeneous MIL benchmarks, where it attains the highest primary classification point estimate on three tasks and remains competitive on the other two. Controlled duplication experiments show targeted robustness to nuisance multiplicity, and on post-stroke FAC prediction the ordinal variant achieves 75% accuracy and 0.30 MAE versus 45% and 0.70 for mean pooling.

1 INTRODUCTION

Multiple-instance learning (MIL) predicts a bag-level target from a collection of instances when instance-level labels are unavailable (Dietterich et al., 1997; Foulds & Frank, 2010; Carbonneau et al., 2018). In many applications, however, instances within a bag are strongly correlated: neighboring image regions, repeated sensor events, or multiple realizations of the same underlying pattern may carry nearly the same evidence. Repeated sampling can therefore increase influence without adding a comparable amount of new information. Yet multiplicity is not always nuisance variation, since relative prevalence may itself be predictive. MIL aggregation should therefore distinguish *nuisance multiplicity* from *informative prevalence*.

Modern MIL methods use learned instance importance (Ilse et al., 2018; Lu et al., 2021), relation-aware aggregation (Li et al., 2021), and set or transformer architectures (Zaheer et al., 2017; Lee et al., 2019; Shao et al., 2021). Recent work also revisits representation and aggregation (Du et al., 2025; Shao et al., 2026) or explicitly handles redundancy (Yao et al., 2026; Sun et al., 2026). Max pooling is unchanged by exact duplicates, whereas mean and softmax attention can change with relative bag composition; none explicitly separates redundancy from informative prevalence. Determinantal constructions model diversity (Kulesza & Taskar, 2012), including in MIL representations (Qiao et al., 2017; Zhu et al., 2024). DENO instead uses an oppositely evidence-tilted log-determinant contrast directly as signed bag-level evidence. Rather than using the determinant for diversity regularization or selection, it couples this redundancy-aware evidence with a separate mean branch that retains relative prevalence.

We introduce **DENO**, a general signed determinantal evidence pooling operator for correlated MIL. Each instance provides signed evidence e_i together with a similarity representation m_i that describes its relationship to other instances. The key design principle is to model redundancy and prevalence through separate aggregation paths rather than force a single pooling rule to handle both. If several similar instances provide the same evidence, their repeated support should have diminishing influence because they contribute little new information. At the same time, how frequently such evidence occurs

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.

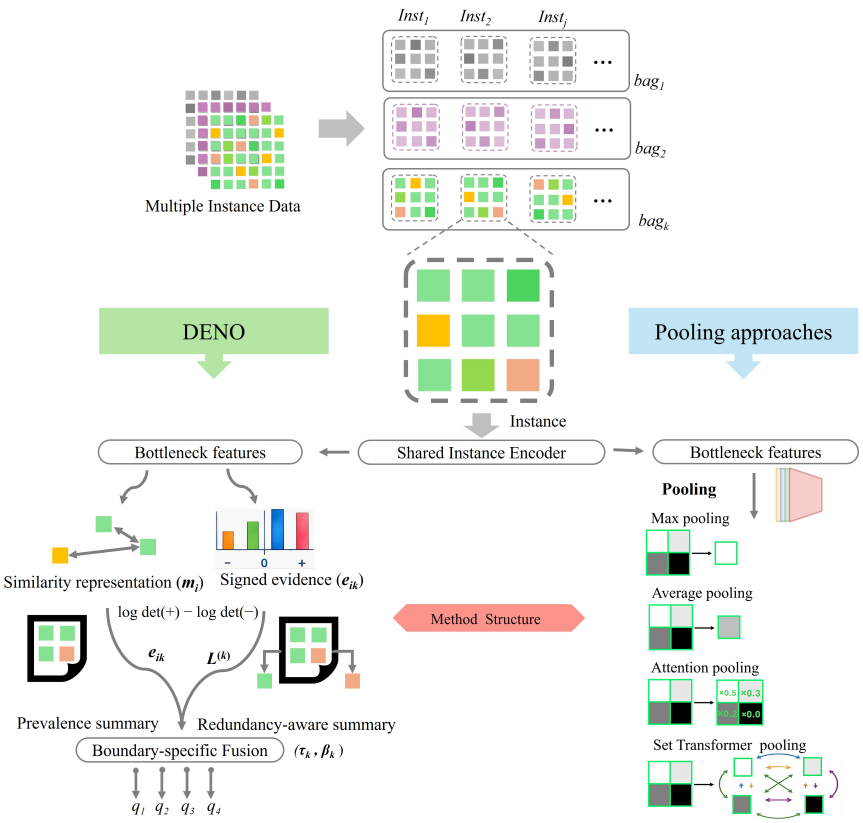


Figure 1: **DENO overview.** Signed log-determinant evidence captures redundancy-aware support, while a separate mean-evidence branch retains relative prevalence.

may itself be predictive and should not be discarded. DENO therefore combines a redundancy-aware determinantal branch, which discounts evidence repeated along similar instance directions, with a mean-evidence branch that retains relative prevalence. The two branches are combined through a learned mixture to produce the bag-level score. Figure 1 illustrates this two-branch design and its boundary-specific extension for ordinal prediction.

This formulation also enables a precise analysis of redundancy. For exact replicas, we show that the determinantal evidence computed over all instances can be reduced exactly to a computation over the distinct underlying/latent sources. As a consequence, uniformly replicating the entire bag leaves both the determinantal and mean-evidence branches, and hence the complete DENO score, unchanged for fixed model parameters. Selectively replicating only some sources, however, can change the score. We then extend the analysis to near-redundant instances, showing that groups of highly similar instances can be approximated by their dominant similarity directions with a bounded error. These results further yield sufficient conditions under which the DENO score, and ultimately the predicted label, remains stable under such approximations. The guarantees characterize the behavior of a fixed trained model and do not imply prediction correctness or generalization.

We evaluate DENO in two complementary ways. First, controlled exact- and near-duplication experiments isolate the effect of nuisance multiplicity and directly test DENO's intended redundancy-aware behavior. Second, we evaluate DENO across five heterogeneous MIL tasks spanning wearable sensing, image bags, and molecular data. DENO achieves the highest primary classification point estimate among the evaluated methods on three of the five tasks and remains competitive on the other two. For ordered bag-level targets, we further develop a boundary-aware ordinal extension based on continuation-ratio modeling (Agresti, 2010; Shi et al., 2023). Applied to post-stroke Functional Ambulation Category prediction (Holden et al., 1984; Zhou et al., 2024), the extension uses the same DENO pooling principle at each ordinal boundary rather than modifying the core operator.

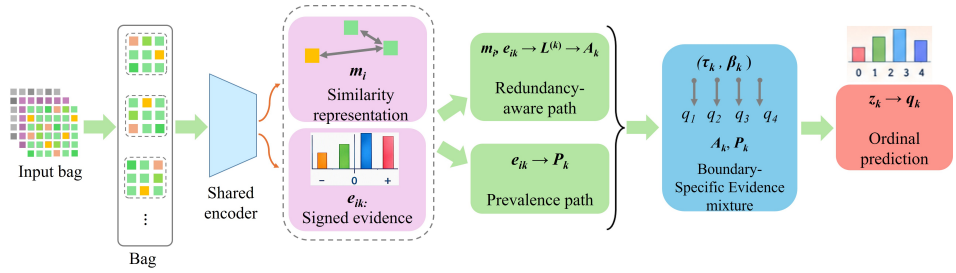


Figure 2: **Boundary-aware DENO architecture.** A shared instance encoder produces similarity representations m_i and signed boundary evidence e_{ik}. The determinantal path produces the redundancy-aware summary A_k, while the separate mean-evidence path retains relative prevalence P_k. Their boundary-specific mixture yields continuation probabilities q_k, which determine the final ordinal class distribution.

Our contributions are fourfold. First, to our knowledge, DENO is the first MIL pooling framework to explicitly separate the two competing roles of instance multiplicity, redundancy and informative prevalence, through complementary aggregation paths. DENO combines signed determinantal pooling, which gives diminishing influence to evidence repeated along correlated similarity directions, with a mean-evidence branch that retains relative prevalence. Second, we provide a rigorous characterization of this mechanism: exact replicas admit an exact reduction to their distinct underlying sources, uniform whole-bag replication leaves the complete DENO score unchanged, and near-redundant instances admit quantitative approximation and prediction-stability guarantees. Third, controlled duplication experiments reveal a striking separation precisely when nuisance multiplicity is introduced: on clean MNIST bags, DENO and mean pooling achieve comparable accuracy (98.25% vs. 98.08%), whereas under IID exact duplication DENO maintains 98.61% accuracy while mean pooling drops to 91.65%; the same pattern persists under near duplication (97.43% vs. 92.71%). Finally, these gains extend beyond controlled settings: DENO achieves the highest primary classification point estimate on three of five heterogeneous MIL tasks, including 75% accuracy and 0.30 MAE on post-stroke FAC prediction (Zhou et al., 2024), compared with 45% accuracy and 0.70 MAE for mean pooling. Together, the theoretical and empirical results show that explicitly distinguishing redundant repetition from informative prevalence can have a substantial practical impact in correlated MIL.

2 DENO: Signed Determinantal Evidence Pooling

DENO is a multiple-instance learning (MIL) pooling operator for bags containing correlated or redundant instances. Its key idea is to distinguish redundancy from prevalence: repeated instances may contribute little new information because they carry similar evidence, yet how frequently a pattern occurs may itself be predictive. DENO therefore uses two complementary aggregation paths: a determinantal branch that discounts evidence repeated along similar instance directions, and a mean-evidence branch that preserves relative prevalence. The two branches are combined to produce the final bag-level evidence score. We first define this generic pooling operator, then establish its behavior under exact and near redundancy, and finally extend it to ordinal prediction using boundary-specific continuation decisions. Figure 2 illustrates the end-to-end ordinal form of DENO.

2.1 Problem Setup and Instance Representation

Let $\mathcal{B} = \{x_i\}_{i=1}^N$ denote a bag of N instances with bag-level target y, and let $h_i = f_\theta(x_i)$ be the encoded representation of instance i. DENO associates each instance with two quantities: $m_i \in \mathbb{R}^{d_m}$ and $e_i \in \mathbb{R}$. The *similarity representation* m_i describes the instance in the feature space used to assess redundancy: instances with nearby m_i are treated as carrying similar information. The scalar *signed evidence* e_i specifies the direction and strength of the instance's support for the bag-level decision. Thus, m_i determines *which instances provide redundant support*, whereas e_i determines *what they support and how strongly*.

The similarity representation may be learned directly from h_i, obtained through auxiliary supervision, or supplied by domain-specific features. Likewise, the evidence may be produced by a learned evidence head. The DENO pooling operator itself is agnostic to these choices and requires only the pairs $\{(m_i, e_i)\}_{i=1}^N$.

2.2 Gaussian Similarity and Degree Normalization

For bandwidth $\ell > 0$, define matrices C and L with elements:

$$C_{ij} = \exp\left(-\frac{\|m_i - m_j\|_2^2}{2\ell^2}\right), \qquad d_i = \sum_{j=1}^{N} C_{ij}, \qquad L_{ij} = \frac{C_{ij}}{\sqrt{d_i d_j}}. \tag{1}$$

Here C is the Gaussian similarity matrix, d_i is the total similarity degree of instance i, and L is the symmetric degree-normalized kernel. Dense neighborhoods of similar instances have larger degree, so normalization reduces their repeated influence without removing any instance. Since the Gaussian kernel is positive semidefinite (PSD), $C \succeq 0$, and the symmetric normalization also gives $L \succeq 0$.

2.3 Signed Determinantal Evidence Pooling

We first define the operator for a generic PSD kernel $G \in \mathbb{R}^{n\times n}$ and evidence vector $v \in \mathbb{R}^n$:

$$D_{\pm}(v) = \operatorname{diag}\left(e^{\pm v_1/4}, \ldots, e^{\pm v_n/4}\right)$$

and

$$\boxed{\mathcal{A}(G, v) = 2[\log\det(I_n + D_+(v)GD_+(v)) - \log\det(I_n + D_-(v)GD_-(v))]\,.} \tag{2}$$

The first log-determinant applies the positive evidence tilt and the second the opposite tilt to the same PSD geometry; thus *signed* refers to their evidence contrast, not to an indefinite kernel. Because $\log(1+\lambda)$ is concave, additional spectral mass along an already strong direction gives a progressively smaller log-determinant increase. Essentially, $\mathcal{A}(G, v)$ aggregates signed instance evidence while giving diminishing influence to evidence repeated along already represented similarity directions.

For a bag, we set

$$G = L, \qquad v = e = (e_1, \ldots, e_N)^\top.$$

The $1/4$ exponent compensates for applying the tilt on both sides of the kernel; together with the outer factor 2, it gives

$$\mathcal{A}(I_n, v) = \sum_{i=1}^{n} v_i, \tag{3}$$

so independent similarity directions recover ordinary additive evidence.

Here we define z as the final bag-level evidence score that combines redundancy-aware determinantal evidence with prevalence-sensitive mean evidence. Because multiplicity may also carry information through relative prevalence, we retain

$$P(e) = \frac{1}{N}\sum_i e_i$$

and define

$$\boxed{z = \beta\frac{\mathcal{A}(L, e)}{\tau} + (1-\beta)P(e),} \qquad \beta \in [0, 1], \quad \tau > 0. \tag{4}$$

Here β balances redundancy-aware and prevalence-sensitive evidence, while τ sets the scale of the determinantal term. The complete predictor is not claimed to be invariant to arbitrary changes in multiplicity.

2.4 Theoretical Guarantees

The following results analyze a fixed trained model. Instance groups are introduced only to characterize redundancy in the theoretical analysis; DENO itself does not form or use such groups during inference.

Exact replicated sources. We first consider the idealized case in which some instances in a bag are exact copies of others. Although the bag contains N instances, suppose there are only $J \leq N$ distinct instances, which we call sources, and every observed instance is an exact replica of one of these sources. We ask whether the determinantal evidence computed from all N instances can be represented exactly using only these J distinct sources. Let $Z \in \{0,1\}^{N\times J}$ map each instance to one source, and let

$$\nu = Z^\top \mathbf{1}_N \in \mathbb{N}^J$$

contain the source multiplicities. Let $S \succeq 0$ be an entrywise-nonnegative source-similarity matrix and $a \in \mathbb{R}^J$ the source evidence. Exact replication gives

$$C = ZSZ^\top, \qquad e = Za.$$

With $q = S\nu > 0$, the source-level degrees induced by similarity and multiplicity, define

$$H_{\text{eff}}(\nu) = \operatorname{diag}\Big(\sqrt{\nu/q}\Big)\, S \operatorname{diag}\Big(\sqrt{\nu/q}\Big),$$

where division and square root are elementwise.

Theorem 1 (Exact Replicated-Source Reduction)**.**

$$\boxed{\mathcal{A}(L, Za) = \mathcal{A}(H_{\text{eff}}(\nu), a).}$$

Thus the N-instance determinant reduces exactly, not approximately, to a J-source determinant. Selective replication may change H_{eff}, whereas uniform replication does not:

$$H_{\text{eff}}(c\nu) = H_{\text{eff}}(\nu), \qquad c \in \mathbb{N}_{>0}.$$

The mean branch is likewise unchanged by uniform whole-bag replication, so the complete DENO score is invariant to this operation for fixed model parameters. In the idealized kernel-separated case $S = I_J$, $\mathcal{A}(L, Za) = \sum_j a_j$ regardless of source multiplicities.

Near-redundant sources. For near-duplicate rather than exact-duplicate instances, we approximate each group of highly similar instances by its dominant similarity direction and measure how much information is lost by discarding the remaining directions. Partition the bag into groups $G_1, \ldots, G_J$. Let $L_{jj} = L[G_j, G_j]$ have eigenvalues $\lambda_{j1} \geq \lambda_{j2} \geq \cdots \geq 0$, and choose a unit nonnegative leading eigenvector u_j, zero-extended outside G_j. Define

$$U = [u_1, \ldots, u_J], \qquad H_{\text{comp}} = U^\top L U, \qquad R_{\text{spec}} = \sum_j \sum_{r=2}^{|G_j|} \lambda_{jr}.$$

Here H_{comp} retains one dominant direction per group and their cross-group interactions, while R_{spec} measures the remaining within-group similarity structure.

Let a_j be representative evidence for G_j. If $|e_i - a_j| \leq \delta_j$ for $i \in G_j$, and $B \geq \max\{\|e\|_\infty, \|a\|_\infty\}$, then:

Theorem 2 (Spectral Stability under Near-Redundancy)**.**

$$\boxed{|\mathcal{A}(L, e) - \mathcal{A}(H_{\text{comp}}, a)| \leq 2e^{B/2} \left[R_{\text{spec}} + \sum_j \delta_j \operatorname{tr}(L_{jj}) \right].} \tag{5}$$

The two terms measure residual within-group geometry and within-group evidence variation. Hence the compressed determinant remains close to the full one when each group is nearly rank one and its members carry similar evidence. For Gaussian kernels, the Appendix A.3 relates these quantities to group radius, bandwidth, and normalized similarity-mass heterogeneity.

Let

$$\mathcal{A}^\star = \mathcal{A}(H_{\text{comp}}, a), \qquad P^\star = \frac{1}{N} \sum_j |G_j| a_j,$$

and let $z^\star$ denote Eq. equation 4 evaluated with $\mathcal{A}^\star$ and $P^\star$. If

$$|\mathcal{A}(L, e) - \mathcal{A}^\star| \le \varepsilon_A, \qquad |P(e) - P^\star| \le \varepsilon_P,$$

then

$$|z - z^\star| \le \frac{\beta}{\tau}\varepsilon_A + (1 - \beta)\varepsilon_P. \tag{6}$$

The Appendix A.4 gives the corresponding ordinal prediction-stability certificate. These guarantees concern perturbations of a fixed trained model, not prediction correctness or generalization.

2.5 Boundary-Aware Ordinal Variant

For an ordered target $y \in \{1, \dots, K+1\}$, we formulate prediction as a sequence of K continuation decisions. Starting from the lowest level, boundary k asks whether a sample that has reached level k should advance beyond it, i.e., $q_k = P(y > k \mid y \ge k, B)$ Thus, a sample with label $y = c$ passes boundaries $1, \dots, c-1$ and stops at boundary c (unless $c = K+1$, in which case it passes all K boundaries). For example, with five ordered levels, a sample with $y = 3$ passes boundaries 1 and 2 and stops at boundary 3.

DENO models each continuation decision using boundary-specific instance evidence. Instance i produces signed evidence e_{ik} for advancing beyond boundary k, with

$$e_{\cdot k} = (e_{1k}, \dots, e_{Nk})^\top.$$

The similarity representation is shared across boundaries, while each boundary has a trainable mixture weight β_k, determinant scale τ_k, and bandwidth ℓ_k. A common scalar initialization is used for all boundaries, after which these parameters are optimized independently:

$$\beta_k = \mathrm{sigmoid}(b_k), \qquad \tau_k = \mathrm{softplus}(t_k), \qquad \ell_k = \mathrm{softplus}(r_k). \tag{7}$$

Initialization details and positive numerical floors are given in the Appendix B.3.

For each boundary,

$$C_{ij}^{(k)} = \exp\left(-\frac{\|m_i - m_j\|_2^2}{2\ell_k^2}\right), \quad d_i^{(k)} = \sum_j C_{ij}^{(k)}, \quad L_{ij}^{(k)} = \frac{C_{ij}^{(k)}}{\sqrt{d_i^{(k)} d_j^{(k)}}}, \tag{8}$$

and

$$z_k = \beta_k \frac{\mathcal{A}(L^{(k)}, e_{\cdot k})}{\tau_k} + (1 - \beta_k)\frac{1}{N}\sum_i e_{ik}. \tag{9}$$

With

$$q_k = \mathrm{sigmoid}(z_k) = P(y > k \mid y \ge k, \mathcal{B}),$$

the continuation-ratio posterior is

$$p(y = c \mid \mathcal{B}) = \left(\prod_{j<c} q_j\right)(1 - q_c), \quad c \le K, \qquad p(y = K+1 \mid \mathcal{B}) = \prod_{j=1}^{K} q_j. \tag{10}$$

Thus class c is obtained by passing all preceding boundaries and stopping at c, while the highest class passes all K boundaries. Only samples with $y \ge k$ contribute to the loss at boundary k.

3 Experiments

We evaluate DENO on five heterogeneous MIL benchmarks, including the standard MNIST-Bags task, together with a separate controlled-MNIST study of exact and near-duplicate nuisance multiplicity and numerical tests of the theoretical results in Sec. 2.4. We also ablate the main components of DENO.

Table 1: Main predictive results. FAC reports five-seed probability ensemble / mean independent-seed performance; other tasks report five-seed means. Higher is better except MAE.

	FAC		Stanford FoG		MNIST-Bags		MUSK2		Mutagenesis	
Method	Acc. (%)↑	MAE↓	Bal.Acc. (%)↑	AUROC↑	Acc.(%) ↑	AUROC↑	Bal.Acc. (%)↑	AUROC↑	Bal.Acc. (%)↑	AUROC↑
Mean	45.0/45.0	0.70/0.69	71.84	0.75	97.58	0.9960	80.50	0.91	75.99	0.86
Max	40.0/32.0	0.70/0.77	71.27	0.74	97.31	0.9955	80.79	0.88	57.15	0.78
LSE	45.0/39.0	0.65/0.65	68.97	0.72	98.02	0.9979	**83.66**	**0.92**	65.28	0.79
Attention MIL	40.0/35.0	0.60/0.76	71.05	0.74	97.86	0.9973	82.48	0.90	74.21	0.84
Gated Attention	35.0/28.0	0.70/0.77	72.06	0.76	97.79	0.9970	81.50	0.89	74.62	0.84
DeepSets	45.0/43.0	0.65/0.68	71.41	0.76	97.44	0.9958	79.12	0.86	73.14	0.87
Set Transformer	45.0/38.0	0.60/0.77	71.61	0.76	97.68	0.9967	80.57	0.88	**82.34**	**0.89**
DENO	**75.0/59.0**	**0.30/0.49**	**73.64**	**0.78**	**98.14**	**0.9981**	83.04	0.91	80.32	0.87

3.1 Experimental Setup

Datasets. We evaluate on five MIL tasks: FAC and Stanford FoG (O'Day et al., 2022) for wearable sensing, MNIST-Bags (Ilse et al., 2018) for image-based MIL, and MUSK2 (Dietterich et al., 1997) and Mutagenesis-Chains for molecular MIL. FAC treats each clinical visit as a bag of walking strides and uses nested patient-level LOSO evaluation. Stanford FoG treats the current IMU window and its seven preceding windows from the same walk as a causal bag and uses subject-level LOSO. MNIST-Bags is the standard existential digit-9 task, while MUSK2 and Mutagenesis-Chains use molecule-level bags. Dataset statistics, preprocessing, and split details are given in B.1.

Separately, controlled MNIST uses six latent sources per bag to isolate nuisance multiplicity: positive bags contain four target and two non-target sources, and negative bags contain two target and four non-target sources. Exact and near-duplicate conditions vary source multiplicity independently of label and source type; full generation details are given in the Appendix B.4.

3.2 FAC Instantiation and Training

The FAC realization uses auxiliary stride-level gait supervision to learn the similarity representation; this application-specific component is not required by generic DENO.

Gait representation. Each bag is one clinical visit and each instance one six-channel foot-IMU stride. A bidirectional GRU encoder produces h_i, and an auxiliary regression head is trained by mean-squared error to predict seven standardized gait quantities: stride length, clearance, stride time, swing time, stance time, stance ratio, and speed. We denote these predictions by $\widehat{g}_i \in \mathbb{R}^7$.

FAC prediction. We use

$$m_i = \widehat{g}_i$$

as the DENO similarity representation. A linear evidence head maps $[h_i; \widehat{g}_i; s_i]$, where $s_i \in \mathbb{R}^2$ is the foot-side indicator, to four boundary evidences $e_{i1}, \ldots, e_{i4}$ for the five FAC levels. The evidence head and $\{\beta_k, \tau_k, \ell_k\}_{k=1}^4$ are trained from visit-level FAC supervision. In the reported configuration, the final encoder projection and gait-regression head are also fine-tuned while earlier auxiliary layers remain fixed, and the gait-regression MSE is retained during FAC training.

For boundary balancing, let $n_k^+ = \#\{y > k\}$ count training samples that pass boundary k, and $n_k^- = \#\{y = k\}$ those that stop there. When both are nonzero,

$$w_k^+ = \frac{n_k^+ + n_k^-}{2n_k^+}, \qquad w_k^- = \frac{n_k^+ + n_k^-}{2n_k^-};$$

if only one group is present, it receives unit weight. The visit-level continuation loss is

$$\mathcal{L}_{\mathrm{CR}} = \sum_{k<y} w_k^+ \,\mathrm{softplus}(-z_k) + \mathbf{1}\{y \le K\} w_y^- \,\mathrm{softplus}(z_y). \tag{11}$$

FAC supervision remains visit-level; no stride-level FAC classification target is used. The training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CR}} + \lambda_e \mathcal{R}_e + \lambda_{\beta,\tau} \mathcal{R}_{\beta,\tau} + \lambda_\ell \mathcal{R}_\ell + \lambda_g \mathcal{L}_{\mathrm{gait}}, \tag{12}$$

Table 2: Controlled MNIST under nuisance multiplicity with matched train/test conditions.

Condition	Method	Accuracy (%) ↑	AUROC ↑	NLL ↓
Clean	Mean	98.08 ± 0.41	0.9960 ± 0.0004	0.0583 ± 0.0057
	Attention	$\mathbf{98.41 \pm 0.30}$	$\mathbf{0.9988 \pm 0.0004}$	$\mathbf{0.0471 \pm 0.0065}$
	DENO	98.25 ± 0.74	0.9980 ± 0.0010	0.0576 ± 0.0171
IID exact duplicates	Mean	91.65 ± 0.84	0.9761 ± 0.0044	0.1985 ± 0.0179
	Attention	91.57 ± 0.72	0.9755 ± 0.0030	0.2033 ± 0.0121
	DENO	$\mathbf{98.61 \pm 0.44}$	$\mathbf{0.9984 \pm 0.0011}$	$\mathbf{0.0494 \pm 0.0184}$
Near duplicates	Mean	92.71 ± 0.57	0.9801 ± 0.0018	0.1851 ± 0.0122
	Attention	92.80 ± 0.74	0.9804 ± 0.0027	0.1819 ± 0.0151
	DENO	$\mathbf{97.43 \pm 0.59}$	$\mathbf{0.9967 \pm 0.0012}$	$\mathbf{0.0743 \pm 0.0132}$

where $\mathcal{R}_e$ controls evidence magnitude, $\mathcal{R}_{\beta,\tau}$ regularizes the mixture weights and determinant scales toward their initialization, $\mathcal{R}_\ell$ analogously regularizes the bandwidths, and $\mathcal{L}_{\text{gait}}$ retains the auxiliary gait regression objective. Exact definitions, coefficients, numerical safeguards are given in Appendix.

3.3 Main Predictive Results

Table 1 compares DENO with standard pooling and learned set-aggregation baselines across five benchmark tasks. For FAC, we report both the five-seed probability ensemble and the mean performance across the five independently trained seeds; for the other tasks, we report the five-seed mean. DENO achieves the highest primary classification result on three of the five datasets. The largest improvement is on FAC, where the DENO ensemble achieves 75% accuracy and 0.30 MAE, compared with 45% accuracy and 0.70 MAE for mean pooling. The improvement remains clear without ensembling: averaged across individual seeds, DENO increases accuracy from 45% to 59% and reduces MAE from 0.69 to 0.49.

DENO also achieves the highest balanced accuracy and AUROC on FoG (73.64% and 0.78) and the highest accuracy and AUROC on MNIST-Bags (98.14% and 0.9981). The differences on MNIST-Bags are small because several methods already achieve near-ceiling performance, while the separation on FoG is more pronounced, though still smaller than on FAC. Results on the molecular datasets are mixed: LSE performs best on MUSK2, with DENO close behind, while Set Transformer performs best on Mutagenesis. Overall, DENO performs strongly across the five benchmarks.

3.4 Controlled Nuisance Multiplicity

The benchmark comparison assesses performance across heterogeneous MIL tasks but does not isolate the effect of repeated evidence. We therefore use the controlled MNIST task from Sec. 3.1, where source composition determines the class while multiplicity is independent of label and source type. Table 2 compares Mean and DENO on clean bags and under exact and near-duplicate nuisance multiplicity.

On clean bags, all three methods perform strongly, with Attention achieving slightly higher accuracy. Their behavior separates once nuisance multiplicity is introduced. Under IID exact duplication, DENO reaches $98.61 \pm 0.44\%$ accuracy, compared with $91.65 \pm 0.84\%$ for Mean and $91.57 \pm 0.72\%$ for Attention; near duplicates show the same pattern ($97.43 \pm 0.59\%$ versus $92.71 \pm 0.57\%$ and $92.80 \pm 0.74\%$, respectively).

Because source multiplicities are sampled independently, these perturbations do not instantiate the uniform whole-bag replication covered by the exact invariance result. Instead, they test whether DENO can limit the influence of nuisance multiplicity when relative source representation varies within a condition. The much larger separation under duplication than on clean bags is consistent with the intended role of the determinantal branch: attenuating repeated support without discarding informative prevalence.

3.5 Numerical Validation of the Theory

We separately test the identities, bounds, and stability certificates from Sec. 2.4. As summarized in Table 3, the exact replicated-source reduction holds to numerical precision, with median absolute

Table 3: Numerical validation of DENO's theoretical identities, bounds, and stability certificates. Bound ratios are median LHS/RHS; coverage is the fraction of tested perturbations satisfying the sufficient certificate.

Property	Criterion / statistic	Result
Exact replicated-source reduction	absolute reduction error	median $= 2.66\times10^{-14}$; max $= 1.91\times10^{-12}$
Uniform replication	max entrywise $\lvert\Delta H_{\mathrm{eff}}\rvert$; max $\lvert\Delta\mathcal{A}\rvert$	0; 0
Selective replication	$\lVert\Delta H_{\mathrm{eff}}\rVert_F > 10^{-12}$	$1000/1000$ (100%)
Spectral stability	bound violations; median LHS/RHS	$0/2500$; 0.0475
RBF geometric stability	bound violations; median LHS/RHS	$0/2500$; 0.0240
Score perturbation bound	violations of $\lvert z - z^\star\rvert \le \varepsilon_z$	$0/2500$
Binary stability certificate	coverage; changes among certified	56.2%; $0/1405$
Ordinal stability certificate	coverage; changes among certified	74.4%; $0/372$

Table 4: FAC mechanistic ablation under patient-level LOSO. Independent-seed results are mean $\pm$ SD; ensemble results average probabilities across five seeds.

2*Variant	Independent seeds: mean $\pm$ SD			5-seed probability ensemble		
	Acc. (%) $\uparrow$	MAE $\downarrow$	Macro-F1 $\uparrow$	Acc. (%) $\uparrow$	MAE $\downarrow$	Macro-F1 $\uparrow$
Mean pooling (no DENO)	45.0 ± 5.0	0.690 ± 0.055	0.210 ± 0.059	45.0	0.700	0.185
Determinant branch only ($\beta = 1$)	47.0 ± 5.7	0.660 ± 0.103	0.351 ± 0.077	65.0	0.400	0.424
No degree normalization ($L = C$)	52.0 ± 13.5	0.610 ± 0.198	0.353 ± 0.104	60.0	0.500	0.382
Shared ℓ, β, τ across boundaries	54.0 ± 8.2	0.590 ± 0.178	$\mathbf{0.379 \pm 0.081}$	70.0	0.350	0.509
Full DENO	$\mathbf{59.0 \pm 10.8}$	$\mathbf{0.490 \pm 0.125}$	0.363 ± 0.092	**75.0**	**0.300**	**0.538**

error 2.66×10^{-14} and maximum error 1.91×10^{-12}. Uniform replication produces no measured change in either H_{eff} or $\mathcal{A}$. Under selective replication, the effective source geometry changes in all 1,000 tested cases, illustrating the non-invariance permitted by the theory.

Across 2,500 near-redundancy cases, we observe no violations of the spectral, RBF, or propagated score bounds; the median LHS/RHS ratios are 0.0475 and 0.0240 for the spectral and RBF bounds. The binary certificate covers 56.2% of these cases (1,405/2,500), while the ordinal certificate covers 74.4% of a separate 500-case experiment (372/500), with no prediction changes among certified cases. These tests validate sufficient stability conditions, not prediction correctness or generalization.

3.6 Mechanistic Ablation

Finally, Table 4 isolates the main design choices in the boundary-aware FAC model. Mean pooling gives 45% ensemble accuracy and 0.70 MAE. The determinant-only variant reaches 65% and 0.40 after ensembling, although its mean individual-seed accuracy is 47% compared with 45% for mean pooling. Full DENO improves both views, reaching 59% mean accuracy across seeds and 75% ensemble accuracy, with corresponding MAE of 0.49 and 0.30.

Removing degree normalization reduces ensemble accuracy to 60% and increases MAE to 0.50, while sharing ℓ, β, and τ across boundaries yields 70% accuracy and 0.35 MAE, compared with 75% and 0.30 for full DENO. Full DENO also gives the best mean accuracy and MAE across seeds, although the shared-parameter variant has slightly higher mean Macro-F1 (0.379 versus 0.363). These results support degree normalization and boundary-specific calibration in the ordinal model.

4 Discussion and Conclusion

DENO separates redundancy-aware determinantal evidence from prevalence-sensitive mean evidence, with an exact replicated-source reduction, near-redundancy stability bounds, and a boundary-aware ordinal variant. Controlled duplication experiments support the intended redundancy-aware behavior, while results across five heterogeneous MIL tasks provide broader empirical evidence for the approach. Exact invariance is limited to uniform whole-bag replication under the exact-replica model, and the stability guarantees concern perturbations of a fixed trained model rather than prediction correctness or generalization. The FAC evaluation contains only ten participants, and exact dense determinants may become costly for substantially larger bags. Overall, the results support treating redundancy and informative prevalence as distinct roles of multiplicity in correlated MIL.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing and for retrieval and discovery of relevant literature. AI-assisted writing was reviewed and revised by the authors, and literature identified with AI assistance was checked against the original sources. We take responsibility for the final content of this work, including all text, claims, citations, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work did not involve new recruitment of human participants or new data collection. The FAC experiments use an existing stroke-rehabilitation dataset collected independently of this study. We use the data only for research on machine-learning methodology and report results at the aggregate level. Our goal is methodological evaluation rather than clinical diagnosis or treatment, and the proposed model should not be interpreted as a clinical decision-making system. External benchmark datasets are used according to their respective research-use and access conditions.

REPRODUCIBILITY STATEMENT

We provide the model definition, training objectives, optimization details, data preprocessing, evaluation protocols, hyperparameter-selection procedure, and random-seed settings in the main paper and appendix. For the FAC experiments, evaluation is performed with patient-level nested leave-one-subject-out splits, with model selection restricted to the corresponding training subjects, followed by retraining and evaluation on the held-out subject. Multiple random seeds and the stated probability-ensemble procedure are used for the reported ensemble results. The appendix additionally specifies the DENO variants, pooling baselines, benchmark construction, controlled redundancy experiments, and assumptions and proofs for the theoretical results. These details are intended to permit reproduction of both the empirical and theoretical analyses.

REFERENCES

Alan Agresti. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, 2 edition, 2010. doi: 10.1002/9780470594001.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77: 329–353, 2018. doi: 10.1016/j.patcog.2017.10.009.

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3.

Zhaolong Du, Shasha Mao, Xuequan Lu, Mengnan Qi, Yimeng Zhang, Jing Gu, and Licheng Jiao. Rethinking multiple-instance learning from feature space to probability space. In *International Conference on Learning Representations*, 2025.

James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010. doi: 10.1017/S026988890999035X.

Maureen K. Holden, Kathleen M. Gill, Michael R. Magliozzi, John Nathan, and Linda Piehl-Baker. Clinical gait assessment in the neurologically impaired: Reliability and meaningfulness. *Physical Therapy*, 64(1):35–40, 1984. doi: 10.1093/ptj/64.1.35.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5:123–286, 2012. doi: 10.1561/2200000044.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021. doi: 10.1109/CVPR46437.2021.01409.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Johanna O'Day, Marissa Lee, Kirsten Seagers, Shannon Hoffman, Ava Jih-Schiff, Łukasz Kidziński, Scott Delp, and Helen Bronte-Stewart. Assessing inertial measurement unit locations for freezing of gait detection and patient preference. *Journal of NeuroEngineering and Rehabilitation*, 19(1): 20, 2022. doi: 10.1186/s12984-022-00992-x.

Maoying Qiao, Liu Liu, Jun Yu, Chang Xu, and Dacheng Tao. Diversified dictionaries for multi-instance learning. *Pattern Recognition*, 64:407–416, 2017. doi: 10.1016/j.patcog.2016.08.026.

Daniel Shao, Joel Runevic, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, Andrew H. Song, and Faisal Mahmood. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In *International Conference on Learning Representations*, 2026.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021.

Xintong Shi, Wenzhi Cao, and Sebastian Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. *Pattern Analysis and Applications*, 26:941–955, 2023. doi: 10.1007/s10044-023-01181-9.

Guangchao Sun, Qiqi Yan, Jinlian Ma, Xiaofang Zhang, Zhenyu Jiang, Tianyu Xia, and Qi Wang. AGDR-MIL: Attention-graded and de-redundant multiple instance learning for whole-slide image classification. *Biomedical Signal Processing and Control*, 117:109626, 2026. doi: 10.1016/j.bspc.2026.109626.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

Zhao Yao, Zhenmi Xie, Mengxin Tian, Guoqing Wu, Yaonan Wang, and Min Liu. GroupMIL: Semantic group based multiple instance learning for whole slide image analysing. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 6993–7000, 2026. doi: 10.24963/ijcai.2026/778.

Amelia Zafra and Sebastián Ventura. G3p-mi: A genetic programming algorithm for multiple instance learning. *Information Sciences*, 180(23):4496–4513, 2010.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3391–3401, 2017.

Lin Zhou, Torsten Rackoll, Lennard Ekrod, Mircea-Gheorghe Balc, Fabian Klostermann, Bert Arnrich, and Alexander H. Nave. Monitoring and visualizing stroke rehabilitation progress using wearable sensors. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–6, 2024. doi: 10.1109/EMBC53108.2024.10782489.

Wenhui Zhu, Xiwen Chen, Peijie Qiu, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. DGR-MIL: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *Computer Vision – ECCV 2024*, pp. 333–351. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72920-1_19.

A PROOFS OF THEORETICAL GUARANTEES

This appendix proves the two redundancy results from Sec. 2.4: (i) Exact Replicated-Source Reduction and (ii) Spectral Stability under Near-Redundancy. We then specialize the second result to DENO's Gaussian RBF similarity and propagate the resulting evidence bounds to score and ordinal prediction stability.

The analysis is pointwise for a fixed trained model and applies separately to each ordinal boundary k. Boundary-balanced training determines how $e_{\cdot k}$, ℓ_k, β_k, and τ_k are learned, but does not change the forward operator analyzed below. We suppress k when unambiguous.

A.1 SETUP AND EXACT REPLICATED-SOURCE REDUCTION

For a bag of N instances, let $e = (e_1, \dots, e_N)^\top$ denote instance evidence and let $C \in \mathbb{R}^{N\times N}$ be symmetric, positive semidefinite, and entrywise nonnegative. Consistent with Eq. equation 1, define

$$d = C\mathbf{1}_N, \qquad \Delta_d = \operatorname{diag}(d), \qquad L = \Delta_d^{-1/2} C \Delta_d^{-1/2},$$

with $d_i > 0$. Thus C is the instance-similarity kernel matrix and L is its symmetric degree-normalized form. The generic results below require only the stated properties of C; Sec. A.3 specializes it to the Gaussian RBF Gram matrix used by DENO.

For any dimension n, PSD kernel $G \in \mathbb{R}^{n\times n}$ and evidence vector $v \in \mathbb{R}^n$, we index the two main-text tilts D_+ and D_- by a branch sign $s \in \{+1, -1\}$ and write

$$D_s(v) = \operatorname{diag}\left(e^{sv_1/4}, \dots, e^{sv_n/4}\right).$$

Here $D_s(v)$ is the evidence-tilt matrix for branch s. The signed determinantal evidence operator is

$$\mathcal{A}(G, v) = 2\left[\log\det(I_n + D_+(v)GD_+(v)) - \log\det(I_n + D_-(v)GD_-(v))\right]. \tag{13}$$

DENO uses $\mathcal{A}(L, e)$. Since $G \succeq 0$, $I_n + D_s(v)GD_s(v) \succ 0$, so both log-determinants are well defined in exact arithmetic.

We first show that an N-instance bag containing exact copies of only J latent sources can be represented by a $J \times J$ source-level kernel.

The assumptions of Theorem 1 are

$$C = ZSZ^\top, \qquad e = Za, \qquad \nu = Z^\top\mathbf{1}_N, \qquad q = S\nu > 0,$$

where $Z \in \{0, 1\}^{N\times J}$ is the source-membership matrix, $\nu \in \mathbb{N}^J$ contains source multiplicities, $S \succeq 0$ is entrywise nonnegative, and $a \in \mathbb{R}^J$ contains source-level evidence. As in the main text, q denotes the source-level degree vector within this subsection. Define

$$H_{\text{eff}}(\nu) = \operatorname{diag}\left(\sqrt{\frac{\nu}{q}}\right) S \operatorname{diag}\left(\sqrt{\frac{\nu}{q}}\right), \tag{14}$$

with elementwise division and square root. Thus $H_{\text{eff}}(\nu)$ is the effective source-level normalized similarity kernel induced jointly by source similarities and source multiplicities.

Proof of Theorem 1. The argument collapses the $N \times N$ determinant to source level using the shared degree and evidence tilt of exact replicas, followed by Sylvester's determinant theorem,

$$\det(I + AB) = \det(I + BA).$$

Every replica of source j has similarity degree

$$q_j = \sum_{\ell=1}^{J} S_{j\ell}\nu_\ell.$$

Thus, with $\Delta_q = \operatorname{diag}(q)$,

$$L = Z\Delta_q^{-1/2} S \Delta_q^{-1/2} Z^\top.$$

For branch $s \in \{+1, -1\}$, define

$$Q_s = \operatorname{diag}\left(e^{sa_1/4}, \ldots, e^{sa_J/4}\right),$$

the source-level evidence-tilt matrix. Since $D_s(Za)Z = ZQ_s$,

$$D_s(Za)LD_s(Za) = ZB_sZ^\top, \qquad B_s = Q_s\Delta_q^{-1/2}S\Delta_q^{-1/2}Q_s.$$

Thus B_s is the tilted normalized source kernel before multiplicities are folded into the determinant.

Let

$$\Lambda_\nu = \operatorname{diag}(\nu) = Z^\top Z$$

denote the diagonal source-multiplicity matrix. By Sylvester's determinant theorem,

$$\det(I_N + D_s(Za)LD_s(Za)) = \det(I_J + B_s\Lambda_\nu).$$

Applying the theorem again gives

$$\det(I_J + B_s\Lambda_\nu) = \det(I_J + \Lambda_\nu^{1/2}B_s\Lambda_\nu^{1/2}).$$

Because all remaining factors are diagonal,

$$\Lambda_\nu^{1/2}B_s\Lambda_\nu^{1/2} = Q_sH_{\text{eff}}(\nu)Q_s.$$

Applying this identity to both branches of Eq. equation 13 gives

$$\mathcal{A}(L, Za) = \mathcal{A}(H_{\text{eff}}(\nu), a),$$

proving Theorem 1. □

Consequences. Under uniform replication $\nu \mapsto c\nu$ with $c \in \mathbb{N}_{>0}$,

$$q \mapsto cq \quad \Longrightarrow \quad H_{\text{eff}}(c\nu) = H_{\text{eff}}(\nu),$$

so the determinantal branch is exactly invariant to uniform replication of the entire bag. The prevalence term from the main text is

$$P(e) = \frac{\sum_j \nu_j a_j}{\sum_j \nu_j},$$

and is likewise invariant. No invariance to arbitrary selective replication is claimed.

In the idealized kernel-separated case $S = I_J$, $H_{\text{eff}} = I_J$ and

$$\boxed{\mathcal{A}(L, Za) = \sum_{j=1}^{J} a_j,} \tag{15}$$

using

$$2[\log(1 + e^{x/2}) - \log(1 + e^{-x/2})] = x.$$

A.2 Spectral Stability under Near-Redundancy

We now relax exact replication and group instances that are only approximately redundant. Partition the bag into disjoint analysis groups $G_1, \ldots, G_J$ and define their membership matrix

$$(Z_G)_{ij} = \mathbf{1}\{i \in G_j\}.$$

For representative group evidences $a = (a_1, \ldots, a_J)^\top$, define

$$e^0 = Z_G a, \qquad e_i^0 = a_j \text{ for } i \in G_j. \tag{16}$$

Thus e^0 is the group-constant evidence vector used as the compressed reference.

Let

$$L_{jj} = L[G_j, G_j]$$

have eigenvalues

$$\lambda_{j1} \geq \lambda_{j2} \geq \cdots \geq 0.$$

Because L_{jj} is entrywise nonnegative, a Perron–Frobenius leading eigenvector can be chosen nonnegative. Choose a unit nonnegative leading eigenvector u_j, embedded by zeros outside G_j. Define

$$U = [u_1, \ldots, u_J], \qquad H_{\text{comp}} = U^\top L U,$$

and

$$R_{\text{spec}} = \sum_{j=1}^{J} \sum_{r=2}^{|G_j|} \lambda_{jr}. \tag{17}$$

Here H_{comp} retains one dominant similarity direction per group together with interactions between dominant directions of different groups. Because $L \succeq 0$ and both L and U are entrywise nonnegative, H_{comp} is positive semidefinite and entrywise nonnegative. A nearly rank-one within-group block represents strong redundancy, so R_{spec} measures the similarity structure remaining outside the retained dominant directions.

Assume

$$|e_i - a_j| \leq \delta_j, \qquad i \in G_j,$$

and choose

$$B \geq \max\{\|e\|_\infty, \|a\|_\infty\}. \tag{18}$$

We now prove the bound stated in Theorem 2:

$$|\mathcal{A}(L, e) - \mathcal{A}(H_{\text{comp}}, a)| \leq 2e^{B/2} \left[R_{\text{spec}} + \sum_{j=1}^{J} \delta_j \operatorname{tr}(L_{jj}) \right]. \tag{19}$$

Proof of Theorem 2. We separate the argument into two steps. First, for group-constant evidence e^0, we bound the error caused by discarding all but one similarity direction per group. Second, we bound the additional change from e^0 to the actual evidence e.

For the first step, consider the group-constant reference e^0. For each group choose V_j so that $[u_j, V_j]$ is an orthogonal basis of that group's coordinate subspace. Thus V_j spans the discarded within-group similarity directions. Combining them gives an orthogonal basis $[U, V]$ in which

$$[U, V]^\top L [U, V] = \begin{pmatrix} H_{\text{comp}} & B_0^\top \\ B_0 & R_0 \end{pmatrix} \succeq 0.$$

Here R_0 is the similarity block on the discarded directions, while B_0 contains their coupling to the retained dominant directions.

Because u_j is supported on G_j,

$$\operatorname{tr}(H_{\text{comp}}) = \sum_{j=1}^{J} u_j^\top L u_j = \sum_{j=1}^{J} \lambda_{j1}.$$

Therefore

$$\operatorname{tr}(R_0) = \sum_j [\operatorname{tr}(L_{jj}) - \lambda_{j1}] = R_{\text{spec}}. \tag{20}$$

Because e^0 is constant within each group, its tilt is scalar on both the retained and residual subspaces of each group. Therefore, in the $[U, V]$ basis,

$$[U, V]^\top D_s(e^0) [U, V] = \begin{pmatrix} Q_s & 0 \\ 0 & T_s \end{pmatrix}, \qquad Q_s = \operatorname{diag}(e^{s a_j/4}),$$

where T_s is block diagonal and equals $e^{s a_j/4} I_{|G_j|-1}$ on the residual subspace of group G_j. Hence

$$[U, V]^\top \left(I + D_s(e^0) L D_s(e^0) \right) [U, V] = \begin{pmatrix} I + Q_s H_{\text{comp}} Q_s & Q_s B_0^\top T_s \\ T_s B_0 Q_s & I + T_s R_0 T_s \end{pmatrix}.$$

The Schur complement of the upper-left block isolates the contribution of the discarded similarity directions. Since

$$I + Q_s H_{\text{comp}} Q_s = Q_s(Q_s^{-2} + H_{\text{comp}})Q_s,$$

we have

$$Q_s(I + Q_s H_{\text{comp}} Q_s)^{-1} Q_s = (H_{\text{comp}} + Q_s^{-2})^{-1}.$$

Define the residual Schur-complement matrix

$$\widetilde{R}_s = R_0 - B_0(H_{\text{comp}} + Q_s^{-2})^{-1} B_0^\top.$$

Orthogonal similarity preserves determinants, so

$$\begin{aligned}\log\det(I + D_s(e^0) L D_s(e^0)) = \; & \log\det(I + Q_s H_{\text{comp}} Q_s) \\ & + \log\det(I + T_s \widetilde{R}_s T_s).\end{aligned}$$

Thus $\widetilde{R}_s$ is the residual contribution of the discarded directions after accounting for their coupling to the retained subspace.

Moreover,

$$\begin{pmatrix} H_{\text{comp}} + Q_s^{-2} & B_0^\top \\ B_0 & R_0 \end{pmatrix} = \begin{pmatrix} H_{\text{comp}} & B_0^\top \\ B_0 & R_0 \end{pmatrix} + \begin{pmatrix} Q_s^{-2} & 0 \\ 0 & 0 \end{pmatrix} \succeq 0.$$

Since $H_{\text{comp}} + Q_s^{-2} \succ 0$, its Schur complement is positive semidefinite, and hence

$$0 \preceq \widetilde{R}_s \preceq R_0.$$

The second determinant above is the residual log-determinant contribution of the discarded directions. Denote it by

$$\eta_s = \log\det(I + T_s \widetilde{R}_s T_s).$$

Using

$$0 \le \log\det(I + X) \le \operatorname{tr}(X), \qquad X \succeq 0,$$

together with $0 \preceq \widetilde{R}_s \preceq R_0$ and $T_s^2 \preceq e^{B/2} I$, we obtain

$$0 \le \eta_s \le \operatorname{tr}(T_s \widetilde{R}_s T_s) \le e^{B/2} \operatorname{tr}(R_0) = e^{B/2} R_{\text{spec}}.$$

Therefore

$$\left|\mathcal{A}(L, e^0) - \mathcal{A}(H_{\text{comp}}, a)\right| \le 2 e^{B/2} R_{\text{spec}}. \tag{21}$$

For the second step, we bound the evidence perturbation from e^0 to e. Consider the interpolation

$$e(t) = e^0 + t(e - e^0), \qquad 0 \le t \le 1,$$

for which equation 18 implies $|e_i(t)| \le B$. To track the log-determinant along this path, define the corresponding tilted kernel

$$M_s(t) = D_s(e(t)) L D_s(e(t)).$$

Let E_i denote the diagonal matrix with a one in entry (i, i) and zeros elsewhere. Since

$$d\log\det(I + M) = \operatorname{tr}\left((I + M)^{-1} dM\right)$$

and

$$\frac{\partial D_s}{\partial e_i} = \frac{s}{4} E_i D_s,$$

we have

$$\frac{\partial M_s}{\partial e_i} = \frac{s}{4}(E_i M_s + M_s E_i),$$

and therefore, writing $M_\pm = M_\pm(t)$ along the path,

$$\frac{\partial \mathcal{A}}{\partial e_i} = [M_+(I + M_+)^{-1}]_{ii} + [M_-(I + M_-)^{-1}]_{ii}. \tag{22}$$

Since

$$0 \preceq M(I + M)^{-1} \preceq M$$

and

$$[M_s]_{ii} = e^{se_i/2} L_{ii} \le e^{B/2} L_{ii},$$

we obtain

$$\sum_{i \in G_j} \frac{\partial \mathcal{A}}{\partial e_i} \le 2e^{B/2} \operatorname{tr}(L_{jj}).$$

The derivatives in Eq. equation 22 are nonnegative, so along the path $e(t)$,

$$\left| \frac{d}{dt} \mathcal{A}(L, e(t)) \right| \le \sum_j \delta_j \sum_{i \in G_j} \frac{\partial \mathcal{A}}{\partial e_i}.$$

Integrating over $t \in [0, 1]$ therefore gives

$$|\mathcal{A}(L, e) - \mathcal{A}(L, e^0)| \le 2e^{B/2} \sum_j \delta_j \operatorname{tr}(L_{jj}).$$

Combining this with equation 21 proves Eq. equation 19, and hence Theorem 2. □

A.3 Gaussian RBF Geometric Specialization

We now connect the abstract spectral residual to the geometry of the Gaussian RBF representation used in the main text. Let m_i denote the similarity representation. The Gaussian RBF Gram (kernel) matrix of these representations is

$$C_{ip} = \exp\left(-\frac{\|m_i - m_p\|_2^2}{2\ell^2} \right), \tag{23}$$

with boundary-specific $\ell = \ell_k$ in the ordinal model.

Suppose group G_j lies within radius r_j of a representative μ_j:

$$\|m_i - \mu_j\|_2 \le r_j, \qquad i \in G_j.$$

Then, for $i, p \in G_j$,

$$C_{ip} \ge \rho_j := e^{-2r_j^2/\ell^2}.$$

Thus ρ_j is a lower bound on every within-group Gaussian similarity.

Define

$$\widehat{d}_i = \sum_{p \in G_j} C_{ip}, \qquad d_i = \sum_{p=1}^{N} C_{ip}, \qquad \alpha_i^2 = \frac{\widehat{d}_i}{d_i}.$$

Hence α_i^2 is the fraction of instance i's total similarity mass that lies within its analysis group. For a Gaussian RBF kernel, $C_{ip} > 0$ in exact arithmetic, so $\widehat{d}_i > 0$, $d_i > 0$, and $\alpha_i > 0$.

Let

$$\bar{\alpha}_j^2 = \max_{i \in G_j} \alpha_i^2, \qquad \underline{\alpha}_j^2 = \min_{i \in G_j} \alpha_i^2, \qquad \omega_j = \bar{\alpha}_j^2 - \underline{\alpha}_j^2.$$

Thus ω_j measures variation in the fraction of each instance's total similarity mass that remains inside its redundancy group; because d_i also contains similarities outside G_j, it can reflect heterogeneous interaction with the rest of the bag.

Write

$$R_{\text{spec},j} = \sum_{r=2}^{|G_j|} \lambda_{jr}, \qquad R_{\text{spec}} = \sum_{j=1}^{J} R_{\text{spec},j}.$$

Proposition 1 (RBF Geometric Stability)**.** *To summarize the effect of group size and the minimum within-group similarity, define*

$$t_j = \frac{|G_j|}{1 + (|G_j| - 1)\rho_j}.$$

This quantity provides a convenient upper-bound factor for the trace of the locally normalized within-group kernel. Then

$$\mathrm{tr}(L_{jj}) \le \bar{\alpha}_j^2 t_j, \tag{24}$$

$$R_{\mathrm{spec},j} \le \bar{\alpha}_j^2 (t_j - 1) + \omega_j, \tag{25}$$

and

$$t_j - 1 \le e^{2r_j^2/\ell^2} - 1, \qquad t_j \le e^{2r_j^2/\ell^2}.$$

Consequently,

$$\boxed{\begin{aligned} &|\mathcal{A}(L, e) - \mathcal{A}(H_{\mathrm{comp}}, a)| \\ &\le 2e^{B/2} \sum_{j=1}^{J} \Big[\bar{\alpha}_j^2 (e^{2r_j^2/\ell^2} - 1) + \omega_j + \delta_j \bar{\alpha}_j^2 e^{2r_j^2/\ell^2} \Big]. \end{aligned}} \tag{26}$$

Proof. The proof compares the globally normalized block L_{jj} with the same within-group kernel normalized using only within-group degrees. Define

$$\widehat{L}_j = \widehat{D}_j^{-1/2} C[G_j, G_j] \widehat{D}_j^{-1/2}, \qquad \widehat{D}_j = \mathrm{diag}(\widehat{d}_i).$$

Since every within-group similarity is at least ρ_j,

$$\widehat{d}_i \ge 1 + (|G_j| - 1)\rho_j.$$

Because $C_{ii} = 1$,

$$\mathrm{tr}(\widehat{L}_j) = \sum_{i \in G_j} \frac{1}{\widehat{d}_i} \le t_j.$$

Let

$$A_j = \mathrm{diag}(\alpha_i : i \in G_j).$$

The diagonal matrix A_j converts local within-group normalization into the global degree normalization used by DENO. Since $\alpha_i > 0$, A_j^{-1} is well defined, and

$$L_{jj} = A_j \widehat{L}_j A_j.$$

Hence

$$\mathrm{tr}(L_{jj}) \le \bar{\alpha}_j^2 t_j.$$

Moreover,

$$\widehat{L}_j \widehat{D}_j^{1/2} \mathbf{1} = \widehat{D}_j^{1/2} \mathbf{1},$$

so 1 is an eigenvalue of $\widehat{L}_j$. Since $\widehat{L}_j$ is similar to the row-stochastic matrix $\widehat{D}_j^{-1} C[G_j, G_j]$ and is positive semidefinite, its largest eigenvalue is 1.

Let v_j be a corresponding unit eigenvector and set

$$x = \frac{A_j^{-1} v_j}{\|A_j^{-1} v_j\|_2}.$$

This choice transfers the leading locally normalized direction into the globally normalized coordinates. The Rayleigh quotient gives

$$\lambda_{j1} \ge x^\top L_{jj} x = \frac{1}{\|A_j^{-1} v_j\|_2^2} \ge \underline{\alpha}_j^2.$$

Therefore

$$R_{\mathrm{spec},j} = \mathrm{tr}(L_{jj}) - \lambda_{j1} \le \bar{\alpha}_j^2 (t_j - 1) + \omega_j.$$

Finally,

$$t_j - 1 = \frac{(|G_j| - 1)(1 - \rho_j)}{1 + (|G_j| - 1)\rho_j} \le \frac{1 - \rho_j}{\rho_j}.$$

Using

$$\rho_j = e^{-2r_j^2/\ell^2}$$

gives

$$t_j - 1 \le e^{2r_j^2/\ell^2} - 1, \qquad t_j \le e^{2r_j^2/\ell^2}.$$

Substitution into Theorem 2 proves equation 26. □

Interpretation. The bound avoids an explicit multiplicative dependence on raw duplicate count beyond its effect through the normalized geometric and similarity-mass quantities, including α_i and ω_j. For $r_j/\ell \ll 1$,

$$e^{2r_j^2/\ell^2} - 1 = 2r_j^2/\ell^2 + O(r_j^4/\ell^4),$$

so the radius-dependent geometric term vanishes quadratically for tight groups.

A.4 From Evidence Stability to Prediction Stability

We next propagate determinant stability to the complete DENO score and then to predictions. For fixed learned β and τ, write $\mathcal{A} = \mathcal{A}(L, e)$ and $P = N^{-1}\sum_{i=1}^{N} e_i$. The final DENO score is

$$z = \beta\frac{\mathcal{A}}{\tau} + (1-\beta)P. \tag{27}$$

Using the compressed group-level reference associated with $e^0 = Z_G a$, define

$$\mathcal{A}^\star = \mathcal{A}(H_{\text{comp}}, a), \qquad P^\star = \frac{1}{N}\sum_{j=1}^{J} |G_j| a_j, \tag{28}$$

and

$$z^\star = \beta\frac{\mathcal{A}^\star}{\tau} + (1-\beta)P^\star. \tag{29}$$

Here $\mathcal{A}^\star$ is the compressed reference evidence. Theorem 2 bounds $|\mathcal{A}(L, e) - \mathcal{A}^\star|$.

The within-group evidence assumption also gives

$$\boxed{|P - P^\star| \le \frac{1}{N}\sum_{j=1}^{J} |G_j|\delta_j.} \tag{30}$$

Indeed,

$$|P - P^\star| = \left|\frac{1}{N}\sum_{j}\sum_{i\in G_j}(e_i - a_j)\right| \le \frac{1}{N}\sum_{j}|G_j|\delta_j.$$

More generally, if

$$|\mathcal{A} - \mathcal{A}^\star| \le \varepsilon_A, \qquad |P - P^\star| \le \varepsilon_P,$$

then

$$\boxed{|z - z^\star| \le \frac{\beta}{\tau}\varepsilon_A + (1-\beta)\varepsilon_P.} \tag{31}$$

Thus Theorem 2 and equation 30 provide an explicit score perturbation bound for a fixed trained model; retraining is not covered.

For binary prediction with threshold θ, if

$$|z - z^\star| \le \varepsilon \qquad \text{and} \qquad |z^\star - \theta| > \varepsilon,$$

then the predicted label is unchanged.

For ordinal prediction, continuation boundary k asks whether $y > k$ given $y \ge k$. Apply the scalar score construction above separately to each boundary, giving an original score z_k and compressed-reference score $z_k^\star$. Define

$$q_k = \text{sigmoid}(z_k), \qquad q_k^\star = \text{sigmoid}(z_k^\star),$$

where

$$q_k = P(y > k \mid y \ge k, \mathcal{B}).$$

Writing $p_c := p(y = c \mid \mathcal{B})$ and $p_c^\star$ for the corresponding compressed-reference probability, we have

$$p_c = \left(\prod_{j<c} q_j\right)(1 - q_c), \qquad p_c^\star = \left(\prod_{j<c} q_j^\star\right)(1 - q_c^\star), \qquad c \le K,$$

and

$$p_{K+1} = \prod_{j=1}^{K} q_j, \qquad p^\star_{K+1} = \prod_{j=1}^{K} q^\star_j.$$

If, for each boundary,

$$|z_k - z^\star_k| \le \varepsilon_k,$$

then

$$q^-_k = \mathrm{sigmoid}(z^\star_k - \varepsilon_k), \qquad q^+_k = \mathrm{sigmoid}(z^\star_k + \varepsilon_k)$$

satisfy $q^-_k \le q_k \le q^+_k$, and therefore

$$p^-_c = \left(\prod_{j<c} q^-_j\right)(1 - q^+_c), \qquad p^+_c = \left(\prod_{j<c} q^+_j\right)(1 - q^-_c),$$

with

$$p^-_{K+1} = \prod_j q^-_j, \qquad p^+_{K+1} = \prod_j q^+_j.$$

Suppose the reference probabilities have the unique maximizer

$$y^\star = \arg\max_c p^\star_c.$$

If

$$\boxed{p^-_{y^\star} > \max_{c \ne y^\star} p^+_c,} \tag{32}$$

then every admissible perturbed score vector has the same argmax, since

$$p_{y^\star} \ge p^-_{y^\star} > p^+_c \ge p_c.$$

We call equation 32 a *prediction-stability certificate*. It guarantees prediction invariance within the specified perturbation set; it does not certify correctness.

A.5 Supporting Properties and Numerical Stabilization

For completeness, we record several structural properties used to sanity-check the implementation. For any permutation matrix Π, DENO is permutation invariant,

$$\mathcal{A}(\Pi L \Pi^\top, \Pi e) = \mathcal{A}(L, e),$$

odd in the evidence,

$$\mathcal{A}(L, -e) = -\mathcal{A}(L, e),$$

and coordinate-wise nondecreasing by equation 22. It also satisfies

$$\mathcal{A}([1], [e]) = e, \qquad \mathcal{A}(I_N, e) = \sum_i e_i.$$

These follow from determinant invariance, exchange of D_+ and D_-, and

$$2[\log(1 + e^{x/2}) - \log(1 + e^{-x/2})] = x.$$

Numerical stabilization. In exact arithmetic no jitter is required because, with $D_s = D_s(e)$,

$$I_N + D_s L D_s \succ 0.$$

Adaptive jitter only guards against finite-precision numerical effects. In the implementation, Cholesky evaluation starts with jitter 10^{-10} and, only if needed, multiplies it by 10 up to a maximum of 10^{-4}. If branch s uses jitter $\xi_s \ge 0$, define

$$K_s = I_N + D_s L D_s.$$

Define

$$\Delta_s = \log\det(K_s + \xi_s I) - \log\det(K_s).$$

Since $K_s \succeq I$ and $\xi_s \geq 0$,

$$0 \leq \Delta_s = \log\det(I + \xi_s K_s^{-1}) \leq N \log(1 + \xi_s).$$

Let $\mathcal{A}_{\mathrm{num}}$ denote the signed determinantal evidence computed from the jittered branch log-determinants. Therefore,

$$\begin{aligned} |\mathcal{A}_{\mathrm{num}} - \mathcal{A}| &= 2|\Delta_+ - \Delta_-| \\ &\leq 2\max\{\Delta_+, \Delta_-\} \\ &\leq 2N \log(1 + \xi_{\max}), \end{aligned}$$

where

$$\xi_{\max} = \max\{\xi_+, \xi_-\}.$$

.

For the exact Gaussian RBF kernel, $C_{ii} = 1$, so every similarity degree is at least one and the implementation's floor at the working dtype's smallest positive normal value is inactive in exact arithmetic. The theory concerns the exact Gaussian Gram matrix; finite-precision symmetrization is only a numerical correction and is not itself used to claim positive semidefiniteness.

Finally, G_j, μ_j, r_j, R_{spec}, α_i, and ω_j are analysis quantities used to characterize near-redundancy, not clustering operations performed by DENO at inference time.

Complexity. For N instances, pairwise distances between the d_m-dimensional similarity representations cost $O(N^2 d_m)$ and are shared across boundaries. Constructing the K bandwidth-specific normalized kernels costs $O(KN^2)$, while the two dense log-determinants per boundary give $O(KN^3)$ total time. Kernel memory is $O(KN^2)$ when all boundaries are batched and $O(N^2)$ when evaluated sequentially.

B Experimental and Implementation Details

This section provides the implementation and model-selection details. Within each dataset, methods use the same preprocessing and evaluation partitions. Architecture and optimization hyperparameters are selected from the corresponding training/validation data separately for each aggregation method.

B.1 Dataset and Split Details

The five benchmarks comprise two wearable-sensing, two molecular, and one image-based MIL task; controlled MNIST is analyzed separately.

MUSK2 (Dietterich et al., 1997) contains 102 molecule bags, 6,598 conformations, and 166 features. We use stratified five-fold cross-validation at the molecule level with three-fold inner cross-validation for model selection. **Mutagenesis-Chains** contains 188 molecule bags and 5,349 chain instances with 24 features; we use the ten predefined train/test folds of Procedure 1 from Zafra & Ventura (2010).

MNIST-Bags (Ilse et al., 2018) is the standard existential MIL benchmark with digit 9 as the target. Separately, controlled MNIST uses six latent sources per bag: positive bags contain four target and two non-target sources, and negative bags contain two target and four non-target sources. Each condition has 8,000 training, 1,000 validation, and 2,000 balanced test bags. Clean bags contain one instance per source. Under IID exact duplication, each source independently receives a multiplicity sampled uniformly from $\{1, \ldots, 5\}$. Near duplication uses the same multiplicity distribution and additionally applies independent $\{-1, 0, 1\}$-pixel shifts along each axis and Gaussian pixel noise with standard deviation 0.03. Multiplicity is independent of label and source type and therefore represents nuisance variation.

Stanford FoG (O'Day et al., 2022) provides a naturally correlated wearable-sensing task. Ankle IMU signals from seven participants are resampled to $100\,\mathrm{Hz}$ and segmented into 1.28-s windows with 85% overlap. Each bag contains the current and seven preceding windows from the same walk, with no future information. We treat this causal context as a bag of correlated instances and train the aggregation module against a single bag-level target. A window is FoG-positive when at least

Table 5: Principal architecture and model-selection settings.

Dataset	Architecture and search space / settings
MNIST-Bags	2D CNN with $(c, d) \in \{(32, 32), (64, 64)\}$ and dropout 0.2; learning rate $\in \{3 \times 10^{-4}, 10^{-3}\}$.
Stanford FoG	1D CNN with representation dimension $d = 32$; each bag contains the current window and the seven preceding windows from the same walk; learning rate $\in \{3 \times 10^{-4}, 10^{-3}\}$.
MUSK2	MLP $166 \to 32 \to 16$ or $166 \to 64 \to 32$; learning rate $\in \{10^{-3}, 3 \times 10^{-4}\}$.
Mutagenesis-Chains	MLP $24 \to 32 \to 16$ or $24 \to 64 \to 32$; learning rate $\in \{10^{-3}, 3 \times 10^{-4}\}$.
FAC	BiGRU stride encoder with $(h_{\mathrm{GRU}}, d, \mathrm{dropout}) \in \{(16, 8, 0), (32, 8, 0), (32, 16, 0), (32, 16, 0.1)\}$; nested patient-level model selection.
Controlled MNIST	2D CNN with $(c, d) \in \{(32, 32), (64, 64)\}$ and dropout 0.2; DENO mechanism dimension $d_m = 16$.

30% of its duration overlaps an annotated FoG episode; the bag label is that of the current window. Evaluation uses leave-one-subject-out cross-validation.

FAC uses the post-stroke wearable dataset of Zhou et al. (2024). Ten participants in early neurorehabilitation were measured at two visits 1–2 weeks apart, yielding 20 visits in total. The mean times since stroke were 8.5 ± 2.6 and 16.3 ± 2.8 days at Visits 1 and 2, respectively. Participants contributed on average 162.1 ± 50.6 and 175.8 ± 52.4 walking strides at the two visits. Data were recorded at 120 Hz using wearable IMUs; our analysis uses the foot-mounted sensors. Each visit forms one MIL bag, with each segmented walking stride treated as an instance. Following Sec. 3.2, the stride encoder uses auxiliary stride-level supervision to predict seven gait quantities (stride length, clearance, stride time, swing time, stance time, stance ratio, and speed), whose predictions define the DENO similarity representation m_i. The primary FAC target is available only at the visit level. The five observed FAC levels are modeled with four continuation boundaries. Evaluation uses nested patient-level leave-one-subject-out cross-validation.

Protocols and baselines. Data-dependent preprocessing and model selection use only the corresponding training data. Hyperparameters are tuned independently for each method using its training/validation data, and each method is trained with five independent seeds per evaluation split. Metrics are computed per seed under the complete evaluation protocol and reported as mean $\pm$ sample standard deviation across the five seed-level results. For controlled MNIST, all methods and seeds within a condition use identical bags; condition-specific datasets are generated according to the constructions above.

For FAC, both visits of the outer held-out participant are excluded from auxiliary-model training and selection, normalization, model selection, and early stopping. Each seed is evaluated on the 20 held-out visits pooled across the ten LOSO folds. For the ensemble, the five seed-specific class-probability vectors are averaged per visit before evaluation on the same 20 visits.

Within each dataset, all methods use the same dataset-specific instance encoder and preprocessing before their respective aggregation modules, together with the same evaluation splits and seed set. FAC uses a BiGRU, MUSK2 and Mutagenesis-Chains an MLP, both MNIST tasks a 2D CNN, and FoG a 1D CNN. We compare mean, max, normalized log-sum-exp (LSE) (Wang et al., 2018), attention and gated-attention MIL (Ilse et al., 2018), DeepSets (Zaheer et al., 2017), Set Transformer (Lee et al., 2019), and DENO.

We report accuracy and MAE for FAC, balanced accuracy and AUROC for Stanford FoG and the molecular benchmarks, and accuracy and AUROC for MNIST-Bags.

B.2 Architectures and Model Selection

Table 5 summarizes the principal encoder choices and dataset-specific search spaces. For the CNN entries, c denotes channel width and d the output representation dimension.

For each benchmark, encoder capacity, learning rate, and aggregation-specific hyperparameters are selected from validation data for that method. Baseline configurations are therefore selected independently rather than inherited from DENO.

B.3 FAC Training Details

Each six-channel stride is first projected to h_{GRU} dimensions and passed through LayerNorm, GELU, and a one-layer bidirectional GRU with hidden size h_{GRU} in each direction. The terminal states from the two directions are concatenated and projected to the stride representation h_i. A linear auxiliary head predicts the seven standardized gait quantities $\widehat{g}_i \in \mathbb{R}^7$.

For FAC prediction, the predicted gait vector is used as the similarity representation,

$$m_i = \widehat{g}_i.$$

A linear evidence head maps $[h_i; \widehat{g}_i; s_i]$, where s_i denotes foot side, to the four continuation-boundary evidences. Model selection is nested within each outer patient holdout. Both visits of the held-out participant are excluded from normalization, auxiliary-model selection, FAC model selection, and early stopping.

FAC training uses AdamW with weight decay 10^{-4}, at most 300 epochs, patience 30, and minimum improvement 10^{-5}. Results use five training seeds, $\{0, 1, 2, 3, 4\}$. In the reported configuration, the recurrent layers of the pretrained gait encoder remain fixed, while its final output projection and gait-regression head are fine-tuned together with the evidence head and DENO parameters. These trainable parameters use learning rate 5×10^{-4}.

DENO initialization is selected on the inner validation data. The candidate bandwidths are

$$\ell_0 \in \{0.5, 1, 2\},$$

with $\tau_0 = 1$; the reported fine-tuning configuration uses $\beta_0 = 0.5$. The same initialization is used for the four continuation boundaries, after which β_k, ℓ_k, and τ_k are learned separately for each boundary. Here b_k, r_k, and t_k are unconstrained trainable scalars:

$$\beta_k = \mathrm{sigmoid}(b_k), \qquad \ell_k = \mathrm{softplus}(r_k), \qquad \tau_k = \mathrm{softplus}(t_k),$$

with learned bandwidths and determinant scales numerically lower-bounded by 10^{-6}.

For a visit with N strides and K continuation boundaries, the evidence regularizer in Eq. equation 12 is

$$\mathcal{R}_e = \frac{1}{NK} \sum_{i=1}^{N} \sum_{k=1}^{K} e_{ik}^2.$$

Let $\beta_k^{(0)}$, $\tau_k^{(0)}$, and $\ell_k^{(0)}$ denote the selected initial values, and define

$$\bar{\beta}_k = \mathrm{clip}(\beta_k, 10^{-6}, 1 - 10^{-6}), \qquad \bar{\beta}_k^{(0)} = \mathrm{clip}(\beta_k^{(0)}, 10^{-6}, 1 - 10^{-6}).$$

The mixture/scale regularizer is

$$\mathcal{R}_{\beta,\tau} = \frac{1}{K} \sum_{k=1}^{K} \left[\mathrm{logit}(\bar{\beta}_k) - \mathrm{logit}(\bar{\beta}_k^{(0)}) \right]^2 + \frac{1}{K} \sum_{k=1}^{K} \left[\log \tau_k - \log \tau_k^{(0)} \right]^2,$$

and the bandwidth regularizer is

$$\mathcal{R}_\ell = \frac{1}{K} \sum_{k=1}^{K} \left[\log \ell_k - \log \ell_k^{(0)} \right]^2.$$

The retained auxiliary gait loss is the mean squared error in standardized gait space,

$$\mathcal{L}_{\mathrm{gait}} = \mathrm{mean}_{i,d} \left(\widehat{g}_{id}^{(z)} - g_{id}^{(z)} \right)^2.$$

For the reported configuration, the regularization coefficients in Eq. equation 12 are

$$\lambda_e = 10^{-4}, \qquad \lambda_{\beta,\tau} = 10^{-3}, \qquad \lambda_\ell = 10^{-3}, \qquad \lambda_g = 0.1.$$

These coefficients are fixed within each training run; the reported configuration is retained under the nested training/model-selection protocol.

Determinants are evaluated in float64. Cholesky stabilization begins with jitter 10^{-10} and increases it by a factor of ten only when needed, up to 10^{-4}.

B.4 Controlled MNIST

The controlled experiment is separate from the standard MNIST-Bags benchmark. Each bag contains six latent MNIST sources. Digits 5–9 are treated as target sources and digits 0–4 as non-target sources. Positive bags contain four target and two non-target sources, whereas negative bags contain two target and four non-target sources. Each setting contains 8,000 training, 1,000 validation, and 2,000 balanced test bags.

Clean bags contain one instance per source. Under IID exact duplication, each source independently receives a multiplicity sampled uniformly from $\{1, \ldots, 5\}$, independently of the bag label and source type. Near duplicates use the same multiplicity distribution, with independent $\{-1, 0, 1\}$-pixel shifts along each image axis and Gaussian pixel noise with standard deviation 0.03.

The clean, exact-duplicate, and near-duplicate settings are each applied consistently to training, validation, and test bags. Within a given setting, all methods and all five runs use the same generated bags, while model initialization and optimization seeds differ across runs. Thus variation across seeds reflects training rather than a different sampled dataset.

Models are trained with AdamW, learning rate 10^{-3}, and weight decay 10^{-4} for at most 30 epochs. Early stopping uses validation NLL with patience 6 and minimum improvement 10^{-4}. For DENO, $d_m = 16$ and

$$\beta_0 = 0.5, \qquad \ell_0 = 0.5, \qquad \tau_0 = 1,$$

with these parameters subsequently optimized during training.

B.5 Numerical Validation Protocol

The numerical theory tests are separate from the controlled-MNIST experiment. The exact-reduction tests sample entrywise-nonnegative PSD source kernels, source evidence, and positive integer multiplicities in float64. The near-redundancy tests use known analysis groups generated around latent representatives, with representation and evidence perturbations, and compute H_{comp}, R_{spec}, the realized δ_j, and the corresponding theorem bounds directly from each generated bag.

We perform 5,000 exact-reduction evaluations across 1,000 random trials. Uniform- and selective-replication tests use 4,000 and 1,000 cases, respectively. The spectral, Gaussian-RBF, propagated-score, and binary prediction-certificate tests each use 2,500 cases, while the ordinal certificate uses 500 trials.

For inequality tests, a violation is recorded only when

$$\mathrm{LHS} > \mathrm{RHS} + 10^{-9}.$$

For selective replication, the effective source geometry is counted as changed when

$$\left\| H_{\mathrm{eff}}^{\mathrm{after}} - H_{\mathrm{eff}}^{\mathrm{before}} \right\|_F > 10^{-12}.$$

The latter threshold is used only to exclude floating-point roundoff.