Inducing Mixture Consistency in Nonlinear Aerosol Set Encoders by Training

Devon Maywald
Department of Statistics
Texas A&M University
College Station, TX 77843
dmaywald@tamu.edu

Pater A. Bosler
Sandia National Laboratories
Albuquerque, NM 87123
pabosle@sandia.gov

Nicole Riemer
University of Illinois at Urbana-Champaign
Champaign, IL 61820
Nriemer@illinois.edu

Matthew West
University of Illinois at Urbana-Champaign
Champaign, IL 61820
mwest@illinois.edu

Lekha Patel
Sandia National Laboratories
Albuquerque, NM 87123
lpatel@sandia.gov

Abstract

Linear aerosol operations such as mixing and spatial coarse-graining require latent representations that commute with population addition. Scale–shape encoders guarantee this property by design; we test whether it can instead be induced in a Deep Sets encoder whose nonlinear outer map breaks mixture consistency. We train with separate representation-level consistency and mixture-output losses on particle-resolved aerosol populations evaluated in 64-way mixtures. For every tested nontrivial training arity, the consistency loss reduces the median encoder gap from 10^{6}–10^{7} to approximately 10^{-7}, a relative gap at double-precision roundoff, and the learned outer map is affine to within roundoff at sampled convex mixtures. Mixture-output training reduces diagnostic error for both nonlinear and linear-by-construction encoders, which track closely. These results show that training can retrofit empirical mixture consistency into a nonlinear encoder, although architectural linearity remains preferable when an exact guarantee is available.

1 Motivation

Aerosol populations are described by the joint distribution of particle size and composition, the *mixing state* [6]. Particle-resolved models represent it exactly but are too costly at global scale [3]. Modal and sectional models are scalable alternatives but often produce large regional errors [11]. Machine-learned surrogates close some of that gap by learning a compact representation of the mixing state [5, 8]. These models learn a latent representation that may be used to read out diagnostics and perform aerosol operations. Emissions, dilution, advection and coarse-graining are all linear operations that commute directly in latent space only if the encoder is linear on the space of aerosol particles. This property is required of the encoder; AeroMELD [9] imposes it by restricting the encoder to scale–shape form. Linearity is then *architectural*, which excludes every encoder not of that form, including the Deep Sets [10] architecture. We ask whether it can instead be *induced* by

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences.

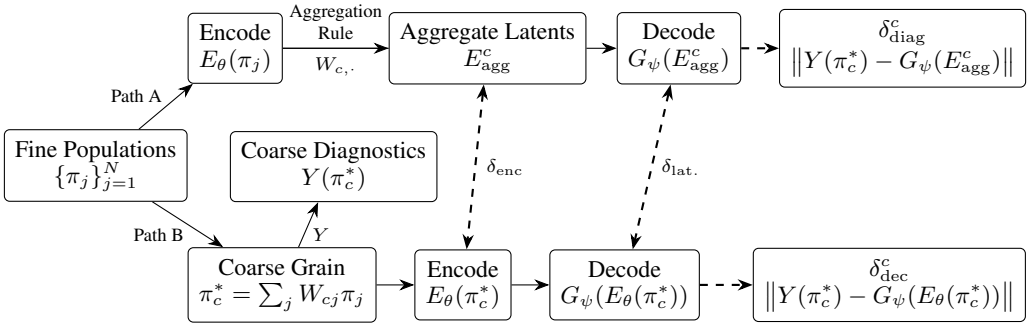

Figure 1: The commuting-diagram test. The upper path (Path A) encodes the fine populations and aggregates in latent space, then calculates coarse diagnostic defects after decoding the aggregated latents. The lower path (Path B) first coarse-grains the populations and then encodes.

the training objective, and contribute such an objective, a decomposition of the diagnostic error that localizes what its consistency term fixes, a diagnostic providing empirical evidence of affineness over a sampled convex hull, and a validating comparison against a linear-by-construction encoder.

2 Populations as weighted sets

A population is a finite weighted set $\pi = \{(a_p, x_p)\}_{p=1}^P$ for a species-mass vector $x_p \in \mathbb{R}^A$ and number concentration $a_p \geq 0$, with P varying by population. Mixing with nonnegative weights w_j acts by concatenation of particle lists with rescaled concentrations, $\sum_j w_j\pi_j = \{(w_j a_p^{(j)}, x_p^{(j)})\}_{j,p}$, so populations $\pi \in \Pi$ form a semimodule; it is with respect to this addition and scaling that an encoder is called linear (1). This analysis is focused on convex combinations of populations, which is of particular interest for aerosol coarse graining

3 Linear and nonlinear set encoders

The reference linear encoder is in scale–shape form. Appending a nonlinear outer map ρ_θ to the pooled representation, in Deep Sets [10] fashion, gives the nonlinear encoder:

$$E^{\mathrm{lin}}_\theta(\pi) = \big(n,\ s(\pi)\big), \qquad E_\theta(\pi) = \big(n,\ \rho_\theta\big(s(\pi)\big)\big), \qquad n = \textstyle\sum_p a_p, \quad s(\pi) = \sum_p \frac{a_p}{n}\,\phi_\theta(x_p). \tag{1}$$

In this case, $\phi_\theta : \mathbb{R}^A \to \mathbb{R}^{L-1}$ is a particle-level multilayer perceptron and s is the number-weighted mean of particle features, the *pooled statistic*. Despite ϕ_θ being nonlinear, E^{lin}_θ is linear on populations, whereas E_θ is not, since ρ_θ does not commute with linear combination. This makes E_θ an ideal controlled test case against E^{lin}_θ since it fails linearity for this isolated reason.

The aggregation rule and mixture consistency. Writing a latent as (n, z), linear operations act in the double-scale coordinates $S = (n, nz)$: for nonnegative weights w_j stacked into a matrix W, aggregation in latent space is ordinary matrix–vector multiplication $S \mapsto WS$, after which $z = (WS)_{2:L}/(WS)_1$, written E_{agg}. This aggregation rule is exact for E^{lin}_θ and applied unchanged to E_θ, where its error is the quantity of interest. The linear aerosol operator employed here takes the block-wise average of spatially distributed populations, an operation we call spatial coarse-graining. Coarse-graining partitions a fine lattice into disjoint equal blocks, so the coarse-cell population is the uniform mean $\pi^*_c = \sum_j W_{cj}\pi_j$. Figure 1 illustrates the two possible pathways from fine populations to coarse-cell diagnostics. Path A encodes each fine population, aggregates, and decodes; Path B forms π^*_c explicitly and encodes it. Path B is unavailable in deployment since a host model never holds the large memory object π^*_c. *Mixture consistency* is agreement between the paths at the representation, $E^c_{\mathrm{agg}} = E_\theta(\pi^*_c)$.

4 Decomposing the output defect

Writing $Y(\pi)$ for the reference deterministic diagnostics and inserting the decoder evaluated on the *exact* coarse encoding gives the following triangle inequality result:

$$\underbrace{\left\|Y(\pi_c^*) - G_\psi(E_{\text{agg}}^c)\right\|}_{\delta_{\text{diag}}} \le \underbrace{\left\|Y(\pi_c^*) - G_\psi(E_\theta(\pi_c^*))\right\|}_{\delta_{\text{dec}}} + \underbrace{\left\|G_\psi(E_\theta(\pi_c^*)) - G_\psi(E_{\text{agg}}^c)\right\|}_{\delta_{\text{lat.}}} \quad (2)$$

per coarse cell c. Since G_ψ is a ReLU network with finite weights, hence Lipschitz on the latent domain considered here [1], the second term is at most $L \cdot \delta_{\text{enc}}$ with $\delta_{\text{enc}} = \|E_\theta(\pi_c^*) - E_{\text{agg}}^c\|$. The *encoder gap* δ_{enc} is the error committed at the representation level and is zero for linear encoders; the *diagnostic gap* δ_{diag} is the readout error of the deployed pathway (Path A Figure 1). The *decoder driven defect* δ_{dec} isolates the error committed by the decoder when given the exact coarse encoding. The *latent mismatch driven defect* $\delta_{\text{lat.}}$ measures the decoder's sensitivity to representation-level mismatch. Because δ_{enc} is measured in latent units, which differ between encoders, we also report the scale-free relative gap $\delta_{\text{enc}}^{\text{rel}} = \delta_{\text{enc}}/\|E_\theta(\pi_c^*)\|$, and likewise $\delta_{\text{diag}}^{\text{rel}} = \delta_{\text{diag}}/\|Y(\pi_c^*)\|$.

The ρ-affineness defect. Those four quantities compare the two paths only at the convex combinations W forms, so a small encoder gap does not by itself establish that the trained nonlinear encoder has become linear. Exact consistency, however, is a strong constraint:

Proposition 1 *Let C be the convex hull of a set of features. If $\rho_\theta\big(ws + (1-w)s'\big) = w\,\rho_\theta(s) + (1-w)\,\rho_\theta(s')$ for all $s, s' \in C$ and $w \in [0,1]$, then ρ_θ preserves every finite convex combination in C and coincides on C with an affine map $s \mapsto As + b$.*

The m-ary case follows from the binary case by induction, writing $\sum_{j\le m} w_j s_j = (1-w_m)\sum_{j<m} \frac{w_j}{1-w_m} s_j + w_m s_m$. A map preserving binary convex combinations on C is the restriction to C of an affine map [7]. The hypothesis is exact consistency for all $s, s' \in C$ and all w; training penalizes violations only approximately and only at finitely many sampled (s, w), so the proposition motivates, rather than implies, affine behavior of the trained map. We therefore quantify the residual affineness of the outer map directly. We draw random convex combinations of held-out populations, evaluate s and $\rho_\theta(s)$ at each, fit $\hat\rho(s) = As + b$ by least squares, and report

$$\delta_\rho = \left\|\rho_\theta(s) - \hat\rho(s)\right\| \Big/ \left\|\rho_\theta(s) - \overline{\rho_\theta(s)}\right\| = \sqrt{1 - R^2}, \quad (3)$$

the fraction of the variation of the outer map that no affine map explains. A value at unit roundoff ($\sim 10^{-16}$) means ρ_θ is affine to working precision at the sampled points; for a ReLU ρ_θ, those points lie on one affine piece of the network.

5 Training modes, architecture, and optimization

Data. Populations are particle-resolved snapshots from the Illinois Data Bank [4], $A = 21$ features per particle, subsampled at a fixed random fraction of 10^{-2} to roughly 1,800 particles each. A pool of 3,000 snapshots is split under a fixed seed into 2,100 training, 150 validation and 750 test populations. All reported defects are measured on the test populations. The diagnostic vector $Y(\pi)$ has dimension 307 (number concentration, bulk species masses and mass fractions, diameter-binned number and speciated mass distributions), computed on the subsampled population, so inputs and targets refer to the same particles. Populations are placed on a 16×16 lattice and averaged into 2×2 cells, a 64-component uniform mixture per cell. Each of 50 trials per run independently draws 256 of the 750 test populations without replacement and places them in random order (200 coarse cells per run; populations recur across trials). Each of 10 seeds re-draws the split, initialization and trials; we report medians over seeds. δ_ρ is fitted on the test populations and read at 512 random convex combinations of them (8-way, Dirichlet($\mathbf{1}$) weights).

Training modes. Three training modes are compared, sharing the same initialization and data partition, so that they differ only by loss function. **Readout-only** freezes E_θ at initialization and trains only the decoder on $L_{\text{single}} = \mathbb{E}_\pi \|G_\psi(E_\theta(\pi)) - Y(\pi)\|^2$; **joint single-population** trains E_θ and G_ψ together on L_{single}; **joint mixture** trains both on the composite training objective

$$L = L_{\text{single}} + \lambda_{\text{out}}\, \mathbb{E}_{\pi,w} \left\|G_\psi(E_{\text{agg}}) - Y(\pi_{\text{mix}})\right\|^2 + \lambda_{\text{enc}}\, \mathbb{E}_{\pi,w} \left\|E_\theta(\pi_{\text{mix}}) - E_{\text{agg}}\right\|^2, \quad (4)$$

Table 1: Mixture arity sweep at $\lambda_{\text{out}} = \lambda_{\text{enc}} = \lambda_{\text{dir}} = 1$. The deployment operator has fan-in 64, so every row extrapolates at least fourfold beyond the arity trained on. Medians over 10 seeds, relative encoder gap $\delta_{\text{enc}}^{\text{rel}}$ and relative diagnostic error in parentheses.

		Encoder gap δ_{enc} ($\delta_{\text{enc}}^{\text{rel}}$)		δ_ρ	Diagnostic gap δ_{diag} ($\delta_{\text{diag}}^{\text{rel}}$)	
Mode	m	nonlinear	linear	nonlinear	nonlinear	linear
Readout-only	–	7.01×10^{6} (1.2×10^{-1})	1.91×10^{-10} (6.4×10^{-18})	2.00×10^{-2}	66.73 (0.446)	62.74 (0.418)
Joint single	–	4.08×10^{7} (3.2×10^{-1})	1.93×10^{-10} (6.4×10^{-18})	1.28×10^{-1}	69.26 (0.461)	65.33 (0.435)
Joint mixture	1	5.68×10^{7} (3.0×10^{-1})	3.62×10^{-10} (1.2×10^{-17})	1.17×10^{-1}	69.71 (0.465)	65.06 (0.431)
	2	1.53×10^{-7} (4.7×10^{-16})	5.13×10^{-10} (1.7×10^{-17})	5.41×10^{-16}	56.66 (0.381)	54.35 (0.366)
	4	1.56×10^{-7} (4.6×10^{-16})	4.85×10^{-10} (1.6×10^{-17})	5.34×10^{-16}	46.96 (0.314)	45.92 (0.307)
	8	1.88×10^{-7} (4.3×10^{-16})	3.59×10^{-10} (1.2×10^{-17})	5.06×10^{-16}	39.74 (0.267)	37.30 (0.251)
	16	2.02×10^{-7} (4.5×10^{-16})	3.39×10^{-10} (1.1×10^{-17})	4.15×10^{-16}	30.95 (0.207)	29.68 (0.199)

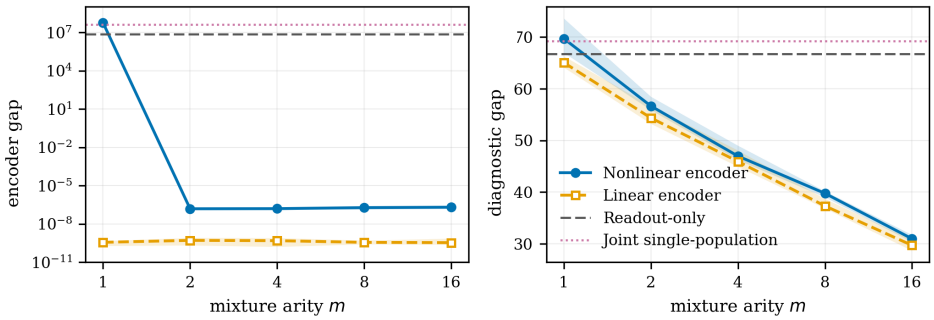

Figure 2: (a) Encoder gap and (b) diagnostic gap across the arity sweep, nonlinear encoder (solid) and linear encoder (dashed); medians over 10 seeds, bands the interquartile range, dashed horizontals the two non-mixture baselines.

with $\pi_{\text{mix}} = \sum_{j=1}^{m} w_j \pi_j$, $\lambda_{\text{out}}, \lambda_{\text{enc}} \geq 0$ and $w \sim \text{Dirichlet}(\lambda_{\text{dir}} \cdot \mathbf{1}_m)$, in which λ_{dir} controls weight sparsity and m the arity. E_{agg} is formed following the aggregation rule 3 using the same generated weights w that forms π_{mix}. The λ_{enc} term enforces mixture consistency directly at the representation level while the λ_{out} term encourages accurate downstream diagnostics. This strategy differs from augmenting the data pool with convex combinations of existing populations, a technique employed by Saleh et al. [9].

Architecture and optimization. ϕ_θ, ρ_θ and G_ψ are multilayer perceptrons of hidden widths $(128, 128, 128)$. Following Saleh et al. [9], the latent shape dimension is 9 so that the decoder maps $\mathbb{R}^{10} \to \mathbb{R}^{307}$. Optimization uses Lion [2] at learning rate 10^{-3}, weight decay 10^{-2}, minibatches of 64, and 20,000 epochs, validating every 25 steps with early stopping after 625 steps without improvement.

6 Mixture arity, and the objective on a second encoder class

Table 1 and Figure 2 sweep the arity m at $\lambda_{\text{out}} = \lambda_{\text{enc}} = \lambda_{\text{dir}} = 1$, applying the training objective to both E_θ and to E_θ^{lin}; Tables 2 and 3 in Appendix A report the Dirichlet and loss-weight ablations.

Mixture consistency is induced at every $m \geq 2$. An unconstrained set encoder does not acquire linearity on its own: training it without ever showing it a mixture makes the encoder gap *worse* than freezing it at initialization (4.08×10^7 against 7.01×10^6), and likewise for $\delta_{\text{lat.}}$ (20.75 against 9.35) and δ_ρ. At $m = 1$ the "mixture" entering the joint mixture training objective is a single rescaled population, a vacuous constraint. At $m = 2$ both δ_{enc} and δ_ρ collapse, the latter to the double-precision floor, holding steady out to $m = 16$, where the sweep stops due to computational budget. The latent mismatch driven defect falls with them from 22 at $m = 1$ to $\sim 3.5 \times 10^{-14}$ for $m \geq 2$. Correspondingly, $\delta_{\text{diag}} = \delta_{\text{dec}}$ to every digit reported meaning the diagnostic gap is purely from the decoder's error. Across all 16 configurations with $\lambda_{\text{enc}} > 0$ and $m \geq 2$ (160 runs, Tables 1, 2 and 3), $\delta_{\text{enc}}^{\text{rel}} \leq 7.3 \times 10^{-16}$ in every run and δ_ρ is at roundoff in 134, whereas with $\lambda_{\text{enc}} = 0$,

$\delta_{\text{enc}} \sim 10^7$–$10^8$ for every λ_{out}. This is consistent with Proposition 1: exact pairwise consistency already pins down an affine outer map, so larger training arities impose no further constraint.

Implications of the collapse to an affine regime. In the 134 runs with δ_ρ at roundoff, ρ_θ is affine to roundoff at every sampled convex mixture. Three consequences follow. First, over the sampled region the outer map ρ_θ is an affine change of coordinates $s \mapsto As + b$, which the decoder's first layer can absorb; training therefore recovers the behavior of E_θ^{lin} rather than creating an encoder that is consistent and more expressive. Second, the evidence is only local: δ_ρ is measured only at sampled convex combinations of held-out populations, so ρ_θ may remain nonlinear elsewhere, and consistency should be checked on the mixtures a given operator produces rather than assumed. Third, the collapse is not always complete: in the other 26 runs, mostly at $m \in \{2, 4\}$, δ_ρ lies between 7×10^{-10} and 2×10^{-3}, yet $\delta_{\text{enc}}^{\text{rel}}$ stays at roundoff. Consistency on the deployment operator therefore does not by itself establish affineness over a wider region, which is why we report both quantities.

The objective improves accuracy when applied to linear-by-construction encoders. Since E_θ^{lin} is linear by construction, its encoder gap sits at 2–5 $\times 10^{-10}$ in every row, including both baselines and for $m = 1$. This is three orders of magnitude below the trained nonlinear encoder only in raw units, which differ between the two latent spaces; relative gaps are $\sim 10^{-17}$ versus $\sim 4.5 \times 10^{-16}$, both at roundoff, so neither encoder is more consistent at working precision. Applying the joint mixture training objective improves accuracy monotonically in m for *both* linear and nonlinear encoders ($0.431 \to 0.199$ relative against $0.465 \to 0.207$). The linear encoder yields slightly lower error at every m, but interquartile ranges overlap at $m \in \{2, 4, 16\}$ and the error is *decoder* driven ($\delta_{\text{lat.}} \sim 10^{-14}$ versus $\delta_{\text{dec}} \sim 10^1$), so this is not evidence that architectural linearity improves accuracy; its advantage is the exact guarantee.

7 Conclusion

Empirical mixture consistency can be learned rather than imposed architecturally, diagnostic improvement comes from mixture-output training, and the nonlinear encoder tracks closely, but does not surpass, the exact linear construction. We found that a representation-level consistency term drives the median encoder gap from 7.0×10^6 to 2×10^{-7}, a relative gap at roundoff in all 160 runs with the term active. A ρ-affineness defect at the double-precision floor provides empirical evidence of affineness over a sampled convex hull of unseen test populations, as anticipated by Proposition 1. Mixture-output training reduces diagnostic error for both nonlinear and linear-by-construction encoders. The evidence comes from one dataset, one operator, one latent dimension and one nonlinear architecture, with the linear comparison covering the arity sweep only; heterogeneous operators (emissions, advection) and other datasets are needed before claiming generality.

References

[1] Trevor Avant and Kristi A. Morgansen. Analytical bounds on the local Lipschitz constants of ReLU networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. URL `https://arxiv.org/abs/2104.14672`. arXiv:2104.14672.

[2] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms, 2023. URL `https://arxiv.org/abs/2302.06675`.

[3] J. H. Curtis, N. Riemer, and M. West. Explicit stochastic advection algorithms for the regional-scale particle-resolved atmospheric aerosol model WRF-PartMC (v1.0). *Geoscientific Model Development*, 17 (22):8399–8420, 2024. doi: 10.5194/gmd-17-8399-2024.

[4] Jessica T. Gasparik, Qing Ye, Jeffrey H. Curtis, Albert A. Presto, Neil M. Donahue, Ryan C. Sullivan, Matthew West, and Nicole Riemer. Data from: Quantifying errors in the aerosol mixing-state index based on limited particle sample size. University of Illinois Urbana-Champaign, Illinois Data Bank, 2020. URL `https://doi.org/10.13012/B2IDB-2774261_V1`.

[5] Michael Hughes, John K. Kodros, Jeffrey R. Pierce, Matthew West, and Nicole Riemer. Machine learning to predict the global distribution of aerosol mixing state metrics. *Atmosphere*, 9(1):15, 2018. doi: 10.3390/atmos9010015.

[6] N. Riemer, A. P. Ault, M. West, R. L. Craig, and J. H. Curtis. Aerosol mixing state: Measurements, modeling, and impacts. *Reviews of Geophysics*, 57(2):187–249, 2019. doi: 10.1029/2018RG000615.

[7] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[8] Ehsan Saleh, Saba Ghaffari, Jeffrey H. Curtis, Lekha Patel, Peter A. Bosler, Nicole Riemer, and Matthew West. Generative modeling of aerosol state representations, 2025. URL `https://arxiv.org/abs/2510.10361`.

[9] Ehsan Saleh, Saba Ghaffari, Wenhan Tang, Jeffrey H. Curtis, Lekha Patel, Peter A. Bosler, Nicole Riemer, and Matthew West. AeroMELD: A linear embedding of aerosol populations for diagnostics and latent dynamics, 2026. URL `https://arxiv.org/abs/2607.11073`.

[10] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets, 2018. URL `https://arxiv.org/abs/1703.06114`.

[11] Z. Zheng, M. West, L. Zhao, P.-L. Ma, X. Liu, and N. Riemer. Quantifying the structural uncertainty of the aerosol mixing state representation in a modal model. *Atmospheric Chemistry and Physics*, 21(23): 17727–17741, 2021. doi: 10.5194/acp-21-17727-2021.

A Training-mode, Dirichlet, and loss-weight ablations

Tables 2 and 3 report the ablations summarized in Section 6. All runs use the nonlinear encoder at $m = 16$ with the data, architecture and optimization of Section 5, and share the same step cap and early-stopping rule.

Table 2: Training-mode comparison, nonlinear encoder, with the Dirichlet concentration λ_{dir} swept inside the joint mixture mode ($\lambda_{\mathrm{out}} = \lambda_{\mathrm{enc}} = 1$, $m = 16$). Medians over 10 seeds, interquartile range in brackets where the spread exceeds the last digit shown. δ_ρ is evaluated on 512 random convex combinations of test populations.

		(a) Representation-level	
Mode	λ_{dir}	δ_{enc}	δ_ρ
Readout-only	–	7.01 [4.49, 7.78] $\times 10^{6}$	2.00 [0.93, 2.89] $\times 10^{-2}$
Joint single	–	4.08 [3.46, 5.04] $\times 10^{7}$	12.8 [9.6, 22.3] $\times 10^{-2}$
Joint mixture	0.1	2.56×10^{-7}	4.55×10^{-16}
	0.316	1.93×10^{-7}	4.93×10^{-16}
	1	2.02×10^{-7}	4.15×10^{-16}
	3.16	2.10×10^{-7}	4.12×10^{-16}
	10	1.68×10^{-7}	5.71×10^{-16}

		(b) Diagnostic-level		
Mode	λ_{dir}	δ_{dec}	$\delta_{\mathrm{lat.}}$	δ_{diag} ($\delta^{\mathrm{rel}}_{\mathrm{diag}}$)
Readout-only	–	65.14 [63.66, 67.27]	9.35 [8.18, 15.65]	66.73 [63.58, 68.78] (0.446)
Joint single	–	64.58 [63.53, 66.66]	20.75 [18.53, 23.68]	69.26 [68.98, 70.45] (0.461)
Joint mixture	0.1	37.50 [36.87, 40.82]	3.23×10^{-14}	37.50 [36.87, 40.82] (0.251)
	0.316	32.13 [31.32, 33.60]	3.37×10^{-14}	32.13 [31.32, 33.60] (0.215)
	1	30.95 [30.35, 31.72]	3.75×10^{-14}	30.95 [30.35, 31.72] (0.207)
	3.16	30.35 [29.03, 31.06]	3.30×10^{-14}	30.35 [29.03, 31.06] (0.203)
	10	31.15 [30.40, 32.65]	3.85×10^{-14}	31.15 [30.40, 32.65] (0.207)

Table 3: Joint mixture loss-weight sweep at $m = 16$, $\lambda_{\mathrm{dir}} = 1$, nonlinear encoder. Rows index the latent-consistency weight λ_{enc}, columns the mixture-output weight λ_{out}. Medians over 10 seeds; interquartile ranges only in panel (c), the only panel with material spread. Baselines: readout-only ($\delta_{\mathrm{enc}} = 7.01 \times 10^{6}$, $\delta_{\mathrm{diag}} = 66.73$, $\delta_{\rho} = 2.00 \times 10^{-2}$) and joint single-population (4.08×10^{7}, 69.26, 1.28×10^{-1}).

	λ_{out} (mixture output)		
λ_{enc}	0	0.1	1
	(a) Encoder gap δ_{enc}		
0	4.10×10^{7}	7.11×10^{7}	8.78×10^{7}
0.01	2.52×10^{-7}	2.51×10^{-7}	2.09×10^{-7}
0.1	2.81×10^{-7}	2.06×10^{-7}	1.98×10^{-7}
1	2.67×10^{-7}	2.37×10^{-7}	2.02×10^{-7}
	(b) Latent mismatch driven defect $\delta_{\mathrm{lat.}}$		
0	11.71	18.04	26.88
0.01	4.76×10^{-14}	2.87×10^{-14}	3.67×10^{-14}
0.1	4.76×10^{-14}	2.76×10^{-14}	3.64×10^{-14}
1	4.30×10^{-14}	2.75×10^{-14}	3.77×10^{-14}
	(c) Diagnostic gap δ_{diag} ($\delta^{\mathrm{rel}}_{\mathrm{diag}}$)		
0	63.59 [62.8, 65.4] (0.425)	32.50 [30.8, 33.0] (0.218)	30.09 [28.9, 30.3] (0.201)
0.01	63.70 [62.2, 65.5] (0.429)	35.74 [34.3, 37.4] (0.238)	31.37 [29.5, 32.0] (0.209)
0.1	64.51 [62.5, 65.8] (0.435)	34.89 [32.4, 37.4] (0.234)	30.53 [30.3, 31.6] (0.206)
1	63.85 [62.5, 65.7] (0.430)	34.24 [33.5, 37.1] (0.229)	30.95 [30.4, 31.7] (0.207)
	(d) ρ-affineness defect δ_{ρ}		
0	1.13×10^{-1}	1.25×10^{-1}	7.71×10^{-2}
0.01	5.05×10^{-16}	5.23×10^{-16}	4.56×10^{-16}
0.1	5.11×10^{-16}	4.85×10^{-16}	4.41×10^{-16}
1	4.86×10^{-16}	5.89×10^{-16}	4.15×10^{-16}