DECLARED GEOMETRY: A GEOMETRIC PRIOR BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

Representations for prediction should discard target-irrelevant input variation while preserving target-relevant information in a predictive geometry. Information-Bottleneck (IB) methods make this compression explicit but do not condition it on the target, a gap the Conditional Entropy Bottleneck (CEB) closes with a label-conditional KL criterion. Yet per-label translations leave the conditional KL unchanged, so collapsed and arbitrarily separated configurations coexist, and under a strict strong data-processing condition strong compression makes an input-independent encoder globally optimal for idealized CEB. We introduce the Geometric Prior Bottleneck (GPB), which declares the geometry in the label-conditional reference and ties the prediction head, making center collapse strictly KL-expensive and, whenever a feasible comparator has lower KL, excluding center-collapsed optima above an explicit compression threshold: OPB classifies from an equal-radius orthogonal frame through anchor energies; EPB regresses through a projection tied to an isometric label axis. We then analyze what the modified objective learns: strong regularization targets the conditional-expectation anchor rather than the input-independent collapse, and under the stated conditions the tied mean readouts remain Bayes-optimal at every compression strength. Across twelve dataset–backbone settings with IB and non-IB baselines, GPB is best or tied-best in eight and never significantly worse than CEB.

1 INTRODUCTION

Two principles recur in predictive representation learning: discard input variation irrelevant to the target, and preserve target-relevant information in a geometry that supports reliable prediction (Tishby et al., 2000; Papyan et al., 2020).

Geometric and contrastive approaches focus on how target information is organized within a representation. Geometric approaches impose this structure through neural-collapse-inspired frames, equidistant or orthogonal constructions, and related mechanisms, acting on features, prototypes, or model weights (Papyan et al., 2020; Deng et al., 2014; Xu et al., 2022; Liang & Davis, 2023; Pai et al., 2022; Luo et al., 2024); contrastive objectives organize representations by pulling same-label points together (van den Oord et al., 2019; Chen et al., 2020; Khosla et al., 2020; Koromilas et al., 2026; Belucci et al., 2025). These methods make the predictive geometry explicit, but their objectives primarily shape the structure needed for prediction. They do not generally include a separate criterion for controlling the target-irrelevant input variation, leaving compression implicit.

Information-bottleneck methods such as VIB, NIB, and DVCCA make compression explicit by bounding the retained input information, through marginal-prior matching, prior-free estimation, and related objectives (Alemi et al., 2019; Kolchinsky et al., 2019b; Abdelaleem et al., 2025). Despite their different formulations, their compression mechanisms are not conditioned on the target label and therefore do not isolate the input variation retained beyond it. CEB addresses this limitation by introducing a label-conditional reference; its compression criterion, the conditional KL between the encoder posterior and that reference, directly targets the residual input information that remains after the label is known (Fischer, 2020).

Yet the conditional-KL certificate does not by itself control the geometry it matches, and under label ambiguity the failure reaches the full objective (Theorem 2.1). First, the certificate is blind to per-label translations of the bottleneck, so collapsed and arbitrarily separated class-conditional

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.

configurations coexist at every certificate value and the KL alone cannot tell which was learned. Second, under a strict strong data-processing condition on the label channel, sufficiently strong compression makes an input-independent encoder globally optimal for the idealized CEB objective, a canonical instance being symmetric label noise.

To address this, we introduce the Geometric Prior Bottleneck (GPB) framework, which places a declared geometry directly in the label-conditional reference: a conditional Gaussian prior whose means satisfy a prescribed separation structure for discrete labels or a prescribed label metric for continuous targets. We instantiate GPB in two forms. For classification, the Orthogonal Prior Bottleneck (OPB) constructs an equal-radius orthogonal frame and predicts using the anchor energies induced by that frame. For regression, the Equidistant Prior Bottleneck (EPB) maps labels onto an isometric axis and predicts through a projection tied to the same axis. In both cases, the same prior means serve as KL targets and prediction anchors, unifying information compression with explicit target geometry. GPB retains the conditional-KL upper-bound formulation while restricting the reference family; the restriction changes what the objective penalizes rather than merely estimating the same compression criterion with a different reference. Unlike an unconstrained reference, the declared geometry then repairs the degeneracy at both levels: it makes posterior-center collapse strictly KL-expensive and, whenever a feasible comparator attains a smaller KL than this collapse cost, it excludes center-collapsed optima above an explicit compression threshold (Theorem 2.2).

Building on this repair, our analysis proceeds in two directions. First, we characterize the predictors the GPB objective itself selects (Theorem 2.3): strong regularization drives the posterior mean to the conditional-expectation anchor, and under the stated conditions the tied mean readouts stay Bayes-optimal at every compression strength. Second, the excess prediction risk of these readouts is bounded by their excess anchor mismatch. The population-optimum statements hold in an idealized function space; they certify the prediction structure of the modified objective rather than superiority over unconstrained references on real benchmarks.

We evaluate GPB across twelve dataset–backbone settings spanning classification and regression, against the plain backbone and the IB, geometric, and contrastive baselines. GPB is best or tied-best in eight settings and never significantly worse than CEB. A controlled study validates the predicted collapse threshold of idealized CEB and the conditional-expectation anchor of OPB; compression sweeps show GPB retaining accuracy in regimes where CEB degrades; and item-wise ablations attribute the gains primarily to the geometric prior and the IB term.

In summary, our contributions are: (i) an analysis of the two-level limitations of the unconstrained conditional reference: certificate-level geometric ambiguity and objective-level information collapse under label ambiguity; (ii) the GPB framework and its OPB/EPB instances that realize this repair, together with sufficient conditions under which the declared prior excludes center-collapsed optima; (iii) a characterization of what the modified objective learns, strong regularization targeting the conditional-expectation anchor with tied mean readouts that remain Bayes-optimal under the stated conditions, together with an excess-risk bound for these readouts; and (iv) extensive experiments validating the advantages of the GPB framework. An extended discussion of related work is given in Appendix A.5, and notation is summarized in Appendix A.1.

2 METHOD

2.1 LIMITATIONS OF UNCONSTRAINED CONDITIONAL REFERENCES

Let $Y - X - Z$ and let the encoder posterior be $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \operatorname{diag}\sigma^2_\phi(x))$. CEB (Fischer, 2020) replaces the marginal reference of VIB by a label-conditional reference $r(z|y)$ and minimizes

$$\mathcal{L}_{\mathrm{CEB}}(q, r, c) = \mathcal{R}(q, c) + \beta\mathcal{K}(q, r), \tag{1}$$

where $\mathcal{R}(q, c) = \mathbb{E}[-\log c(Y|Z)]$ is the prediction risk of the readout c, $z \sim q(z|x)$, and $\mathcal{K}(q, r) = \mathbb{E}_{p(x,y)}\,\mathrm{KL}(q(z|x) \,\|\, r(z|y))$ is the expected conditional KL. Let $q(z|y) = \int q(z|x)\,p(x|y)\,dx$ be the aggregated class-conditional posterior. The expected conditional KL decomposes as the following variational identity (Fischer, 2020):

$$\mathcal{K}(q, r) = I_q(X; Z|Y) + \mathbb{E}_{p(y)}\,\mathrm{KL}\big(q(z|y) \,\|\, r(z|y)\big), \tag{2}$$

since the second term is non-negative, $\mathcal{K}(q, r)$ certifies the residual input information $I(X; Z|Y) \le \mathcal{K}(q, r)$ that CEB minimizes as its compression criterion. Both terms in Eq. (2) are label-local:

each is an average of per-label quantities and neither compares two distinct labels, so the compression criterion alone leaves the inter-label geometry of the class-conditional means underdetermined. Theorem 2.1 makes this precise at two levels.

Theorem 2.1 (Two-level limitations of unconstrained conditional references)**.** (a) Certificate-level geometric ambiguity. *For deterministic labels $Y = f(X)$ and a feasible family closed under per-label translations of the bottleneck space, the conditional KL of Eq. (1) is invariant under per-label translations; at every attainable certificate value a collapsed and an arbitrarily separated realization of the class-conditional means coexist, and the KL alone cannot tell which was learned.*

(b) Objective-level information collapse under label ambiguity. *Let*

$$\bar{\eta} = \sup_{U - X - Y,\, I(X;U) > 0} \frac{I(Y;U)}{I(X;U)} \tag{3}$$

be the strong data-processing coefficient of the label channel, and suppose $\bar{\eta} < 1$, the reference and readout families are sufficiently expressive, and an input-independent posterior is feasible. Then for the idealized nested objective $\mathcal{L}^{\star}_{\mathrm{CEB}}(q) := \min_{r,c}[\mathcal{R}(q,c) + \beta\mathcal{K}(q,r)]$, the input-independent posterior is a global minimizer whenever $\beta \geq \bar{\eta}/(1-\bar{\eta})$, and at strict inequality every global minimizer satisfies $I(X;Z) = 0$. Proofs, definitions, boundary cases, and the binary-symmetric-noise threshold are collected in Appendix A.2.1.

2.2 Declaring geometry in the conditional prior

The GPB framework. GPB restricts the conditional reference to Gaussians

$$p_\theta(z|y) = \mathcal{N}(m_\theta(y), \tau^2 I), \qquad \tau^2 = 1, \tag{4}$$

with fixed prior covariance; the posterior covariance remains the learned diagonal. The anchor means obey a declared geometry fixed before training: a two-sided separation for every pair of distinct discrete labels,

$$\Delta \leq \|m_\theta(y) - m_\theta(y')\| \leq \bar{\Delta}, \tag{5}$$

with declared margin Δ and ceiling $\bar{\Delta}$, or a two-sided label metric condition for continuous labels,

$$\rho_{\min} d_Y(y,y') \leq \|m_\theta(y) - m_\theta(y')\| \leq \rho_{\max} d_Y(y,y'), \tag{6}$$

i.e., a bi-Lipschitz embedding of the label metric d_Y. These conditions define the geometric prior family $\mathcal{G}$. The objective is

$$\mathcal{L}_{\mathrm{GPB}} = \mathbb{E}[\ell(c_\psi(z), Y) + \beta\, \mathrm{KL}(q_\phi(z|X) \| p_\theta(z|Y))], \qquad z \sim q_\phi(z|X), \tag{7}$$

with $\ell \geq 0$ cross-entropy for classification and the chosen regression loss for continuous labels. The anchor means are reused as prediction anchors through a tied readout, so the declared geometry is shared by compression and prediction. The prior remains label-conditional, so the CEB identity of Eq. (2) still holds with the reference restricted to $p_\theta(z|y)$, defining the GPB certificate $\mathcal{K}(w)$; but restricting the reference family changes what the certificate consists of: the mismatch term now encodes a structural preference for the declared geometry. GPB therefore changes what the objective penalizes rather than estimating the same compression criterion with a different reference, and the same β does not by itself correspond to the same conditional mutual information across methods.

Instances: OPB and EPB. For classification, OPB evaluates a prior encoder on all class labels and uses its thin polar factor Q_θ. If U_θ is the resulting all-class matrix, then

$$Q_\theta = U_\theta (U_\theta^\top U_\theta)^{-1/2}, \qquad Q_\theta^\top Q_\theta = I_K, \tag{8}$$

with U_θ of full column rank. Writing $q_{\theta,k}$ for the k-th column of the frame Q_θ, the anchors and their fixed geometry are

$$p_\theta^{\mathrm{OPB}}(z|Y=k) = \mathcal{N}(a q_{\theta,k}, \tau^2 I), \quad (aq_i)^\top (aq_j) = a^2 \mathbf{1}\{i=j\}, \quad \|aq_i - aq_j\| = \sqrt{2}a\ (i \neq j). \tag{9}$$

The anchor-energy scores are the Gaussian Bayes scores

$$s_k(z) = -\frac{\|z - a q_{\theta,k}\|^2}{2\tau^2} + \log \pi_k, \tag{10}$$

All classification experiments use uniform score priors $\pi_k = 1/K$, one of the conditions of the Bayes-ordering guarantee of Theorem 2.3(b). With uniform priors, the scores of Eq. (10) induce an orthogonal classifier whose scale is fixed by the prior; Appendix A.3.2 gives the polar construction, initialization, and covariance specification, and Appendix A.5 the relation to prototype/ETF methods.

For regression, EPB uses a one-dimensional continuous label, min–max normalized using statistics fitted once on the training set as $\tilde{y}$. Let $W \in \mathbb{R}^{d\times 1}$ be a trainable direction with per-step normalization $u = W/\|W\|$; the isometric prior is

$$p_\theta^{\mathrm{EPB}}(z|y) = \mathcal{N}(\rho\tilde{y}\,u, \tau^2 I), \qquad \|m_\theta(y) - m_\theta(y')\| = \rho|\tilde{y} - \tilde{y}'|, \tag{11}$$

so latent distances equal ρ times normalized label distances exactly. Order and magnitude of label differences are encoded with equality, which makes it a member of $\mathcal{G}$ (Eq. (6) with $\rho_{\min} = \rho_{\max} = \rho$). Why this exactness is automatic, its trade-off against equal anchor norms, and the multidimensional-label extension are given in Appendix A.3.2. The prediction head is the tied projection $\hat{\tilde{y}} = u^\top z/\rho$, the regression analogue of the anchor-energy classifier: no free direction or scale in the readout can bypass the prior.

Theorem 2.2 (Declared geometry excludes collapsed optima)**.** *For OPB, any model whose class-conditional centers all coincide pays*

$$\mathcal{K}(w) \geq C_{\mathrm{col}} = \frac{a^2}{2\tau^2}\Big(1 - \sum_{k=1}^{K} \pi_k^2\Big), \tag{12}$$

with arbitrary covariance and within-class variation, equality being attained by the input-independent posterior at the anchor centroid; for EPB the analogous cost is $C_{\mathrm{col}} = \rho^2 \operatorname{Var}(\tilde{Y})/(2\tau^2)$. If a feasible comparator w_0 attains $\mathcal{K}(w_0) < C_{\mathrm{col}}$, then no center-collapsed model is globally optimal once $\beta > \mathcal{R}(w_0)/(C_{\mathrm{col}} - \mathcal{K}(w_0))$; the fitted-model version with an optimization tolerance is stated and proved in Appendix A.2.1.

Theorem 2.2 requires a feasible comparator with $\mathcal{K}(w_0) < C_{\mathrm{col}}$. The natural candidate is the conditional-anchor comparator $w_0 = (a\sum_k \eta_k(x) q_k,\ \tau^2 I)$, whose posterior mean is the expectation of the anchors under the label posterior; its margin is $C_{\mathrm{col}} - \mathcal{K}(w_0) = \frac{a^2}{2\tau^2}\mathbb{E}\|\eta(X) - \pi\|_2^2$, positive exactly when the input carries predictable label structure, with the analogous EPB condition that $\mathbb{E}[\tilde{Y} \mid X]$ be non-constant. Note that Theorem 2.2 implies neither pairwise separation nor prediction accuracy. Together with Theorem 2.1(b), the two objectives admit opposite optimal structures in the same strong-compression regime on a distribution satisfying the respective assumptions: idealized CEB can demand input-independent encoders, while GPB excludes center-collapsed optima.

2.3 What the GPB objective learns

The previous subsections concerned what the unconstrained reference fails to certify and what the declared geometry rules out; here we characterize the predictors the objective of Eq. (7) itself selects. Throughout, write $\mathcal{R}(w) = \mathbb{E}[\ell(c_\psi(Z), Y)]$, $\mathcal{K}(w) = \mathbb{E}\,\mathrm{KL}(q_\phi(z|X)\|p_\theta(z|Y))$, and $\bar{m}(x) = \mathbb{E}[m_\theta(Y) \mid X = x]$ for the conditional-expectation anchor (for OPB, $\bar{m}(x) = aQ_\theta\eta(x)$; for EPB, $\bar{m}(x) = \rho\nu(x)u$).

The conditional KL then decomposes exactly as

$$\mathcal{K}(w) = \delta_m + A_m(w) + V(\Sigma_\phi), \tag{13}$$

with

$$\delta_m = \frac{\mathbb{E}\|m_\theta(Y) - \bar{m}(X)\|_2^2}{2\tau^2}, \qquad A_m(w) = \frac{\mathbb{E}\|\mu_\phi(X) - \bar{m}(X)\|_2^2}{2\tau^2}, \tag{14}$$

δ_m the label-ambiguity constant, $A_m(w)$ the mean-to-anchor term, and $V(\Sigma_\phi)$ the variance-mismatch term, which is non-negative and vanishes exactly at $\Sigma_\phi = \tau^2 I$; the decomposition follows from the standard diagonal-Gaussian KL formula and the conditional bias–variance identity.

Theorem 2.3 (Conditional-anchor learning and predictive structure)**.** (a) Strong-regularization limit. *Suppose the conditional-anchor comparator $(\bar{m}, \tau^2 I)$ is realizable with finite risk R_0, and the fitted model $\widehat{w}$ is within tolerance ε of it on the population objective. Then*

$$A_m(\widehat{w}) + V(\hat{\Sigma}) \leq \frac{R_0 + \varepsilon}{\beta}, \tag{15}$$

so strong regularization drives the posterior mean toward $\bar{m}(X)$ and the covariance toward $\tau^2 I$; when the conditional-expectation anchor varies with the input, this limit is input-dependent rather than input-independent.

(b) Predictive structure of the tied mean readouts. *For OPB at fixed posterior covariance, the conditional objective is strongly convex in the mean, and its unique optimum stays within $O(\beta^{-1})$ of the geometric embedding of the label posterior,*

$$\left\|\mu^\star_\beta(x) - aQ_\theta\eta(x)\right\|_2 \le \frac{\sqrt{2}a}{\beta}, \tag{16}$$

for every $\beta > 0$; the excess classification risk of the optimal mean readout is then at most $\min\{1,\ 2/\beta\}$ by the predictive-consequence bound, with the exact fixed point characterized in Appendix A.2.2; under isotropic posterior noise, uniform score priors, and a fully expressive mean function, the tied mean readout preserves the Bayes ordering $\arg\max_k \eta_k(x)$ for every $\beta > 0$ (Appendix A.2.2): the objective itself never demands mean-path collapse. For EPB under squared loss and a fixed unit direction u,

$$\mu^\star(x) = \rho\,\mathbb{E}[\tilde{Y} \mid X = x]\,u, \tag{17}$$

so the tied mean prediction is the Bayes regressor for every $\beta > 0$; the full conditional KL keeps readout-direction posterior noise at a positive balance below τ^2 where pure mean matching degenerates (exact form in Appendix A.2.2; the GPB-M ablation of Appendix A.4.4).

3 Experiments

The evaluation addresses three core questions, together with three auxiliary diagnostic views: the white-box PGD robustness study (Sec. 3.4), the posterior-geometry diagnostics (Sec. 3.6), and the information-plane estimates (Sec. 3.8).

(RQ1) End-to-end utility. Does GPB improve the operating point across dataset–backbone settings, and which design factor carries the gain? The main results (Sec. 3.2) report the former, and the item-wise ablations (Sec. 3.5) isolate the latter.

(RQ2) Objective selection under label ambiguity. Under a known label channel, does strong regularization make idealized CEB discard input dependence while GPB approaches the conditional-expectation anchors, and how do the mean and variance components affect prediction? The controlled objective-selection study answers the first two questions (Sec. 3.3); the variance component's mechanism is characterized in Theorem 2.3(b) and Appendix A.2.2, and its empirical contribution is isolated in the ablations (Appendix A.4.4).

(RQ3) Design factors and scales. Which design factors and scales determine the operating point? The ablations (Sec. 3.5), the anchor-scale sensitivity (Appendix A.4.6), and the compression sweeps (Sec. 3.7) read the factors and scales.

3.1 Experimental setup

We evaluate twelve dataset–backbone settings over nine public datasets: MNIST, ImageNet-100, Cora, IMDb, AG News, California Housing, STS-B, ZINC-12k, and AgeDB, covering MLP, CNN, GCN, and RNN backbones. All methods use the same data splits, backbone, Adam training, early stopping, and ten seeds per configuration. Stochastic bottleneck methods use reparameterized sampling during training and posterior-mean inference for the main comparisons. The robustness and information-plane studies use sampled representations, as specified in their respective protocols. The baselines are Base, VIB, SVIB, NIB, CEB, supervised β-DVCCA, and the external non-IB baselines NCBD (unit-hypersphere prototypes with a simplex-ETF contrastive optimum) and AdaCap (contrastive loss with a closed-form readout); GPB is OPB on classification rows and EPB on regression rows. The main comparison equalizes the search budget across methods: the plain backbone has no method-specific hyperparameters; every other method family is searched over 30 configurations. The diagnostics in Secs. 3.4, 3.7, and 3.8 read trained configurations on extended grids beyond the selection grid and play no role. For each method and setting, the configuration with the best mean validation metric is selected (accuracy on classification settings, R^2 on regression settings). The reported results are means with standard deviations over seeds, and paired differences use 95% bootstrap confidence intervals. The complete experimental setup is in Appendix A.3.1.

Table 1: Test accuracy (%) on classification tasks and test R^2 on regression tasks (mean over runs); the best entry per dataset is bolded and the second-best underlined. Dashes mark the settings where a method does not apply. The last column reports, per setting, the number of baselines GPB significantly outperforms / is significantly outperformed by, based on paired confidence intervals.

Dataset	Backbone	Base	VIB	SVIB	NIB	CEB	DVCCA	NCBD	AdaCap	GPB	W/L
MNIST	MLP	98.33	98.28	98.15	98.27	<u>98.43</u>	98.24	97.95	–	**98.54**	7/0
MNIST	CNN	99.13	99.28	99.23	99.22	<u>99.30</u>	99.29	99.18	–	**99.32**	4/0
ImageNet-100	MLP	83.12	82.84	83.02	82.35	82.77	82.99	<u>83.64</u>	–	**84.40**	7/0
ImageNet-100	CNN	<u>83.40</u>	82.94	83.11	82.91	82.97	83.21	80.56	–	**86.08**	7/0
Cora	GCN	<u>87.68</u>	86.83	86.46	86.35	86.27	86.90	87.34	–	**88.12**	6/0
IMDb	RNN	84.85	85.50	85.99	86.14	85.55	85.69	<u>86.69</u>	–	**86.84**	6/0
AG News	RNN	**92.94**	92.72	92.86	<u>92.91</u>	92.83	92.78	92.37	–	92.87	2/0
Cal. Housing	MLP	0.7618	**0.7901**	<u>0.7897</u>	0.7862	0.7894	0.7878	–	0.7758	0.7895	2/0
STS-B	RNN	**0.2899**	0.2719	0.2697	0.2417	0.2668	0.2713	–	0.1613	<u>0.2762</u>	2/1
ZINC	GCN	<u>0.6348</u>	0.6339	0.6325	0.6178	0.6257	0.6294	–	**0.6770**	0.6344	3/1
AgeDB	MLP	0.2180	0.3173	0.3385	0.2710	0.3317	0.3180	–	<u>0.3482</u>	**0.3536**	5/0
AgeDB	CNN	0.3914	0.3342	0.3323	0.3337	0.3882	0.3205	–	<u>0.4684</u>	**0.4700**	6/0
Mean rank		4.08	4.83	4.67	6.00	5.00	5.00	5.43	4.00	1.58	–
Best or tied-best		2	1	0	0	0	0	0	1	8	–

3.2 Main results

Table 1 reports, for every baseline and GPB variant, the results of its best validation-selected configuration; the complete standard deviations and paired confidence intervals are in Appendix A.4.1. The ranking summary is dominated by GPB: it has the lowest mean rank and is best or tied-best in $8/12$ settings (ties are exact ties of the reported means). Across the 84 per-setting paired comparisons, GPB is significantly better in 57 of them and significantly worse in only 2 (STS-B versus the plain backbone; ZINC versus AdaCap); under a Holm–Bonferroni adjustment the counts are 49 and 2. The gains are largest where the task has margin to resolve them: on ImageNet-100 (CNN) GPB leads CEB by 3.11 points and on Cora by 1.85, while the saturated settings are flat (MNIST CNN and AG News, where all methods sit within 0.6 points of the best) and the small-data regressions are essentially tied with the best baselines (Housing within 0.0006 R^2 of VIB). Among the external non-IB baselines, AdaCap is competitive on regression (best on ZINC). Across modalities, the same construction improves the operating point where improvement is available, with significant losses confined to the two settings above.

3.3 What strong regularization selects

We now turn to the selection question itself: on a distribution that is predictable yet ambiguous, do the CEB and GPB objectives choose different representations under strong regularization? This is a controlled population-objective experiment, not a generalization benchmark. The input is uniform binary, $X \sim \mathrm{Bernoulli}(1/2)$, and the label follows a binary symmetric channel $Y = X \oplus N$ with $N \sim \mathrm{Bernoulli}(r)$; the encoder is the shared two-dimensional Gaussian family $q_t(z|x) = \mathcal{N}(\tilde{c} + (2x-1)t\tilde{v},\ I_2)$ with $\tilde{c} = \frac{a}{2}(e_0 + e_1)$ and $\tilde{v} = \frac{e_1 - e_0}{\sqrt{2}}$, parameterized by a trainable scalar $t \geq 0$ controlling the input dependence ($t = 0$: input-independent). On the same search space we minimize, numerically, two objectives: the *CEB-oracle*, the idealized CEB with the exact aggregated reference and Bayes readout, and OPB with its declared anchors and tied energy readout ($a = 2$); the protocol, closed forms, and numerical checks are in Appendix A.4.2.

Fig. 1(a) reads the retained input information of the two objectives. The CEB-oracle keeps nearly the full $\ln 2$ nats at weak compression and collapses to zero at the strong data-processing threshold β_{SDPI} marked in the figure, as Theorem 2.1(b) predicts; the measured transition points match the closed-form threshold at every noise rate (Appendix A.4.2). OPB retains less input information already at weak compression (about 0.33 nats), because its fixed-scale tied readout penalizes overconfident predictions under label noise and thus favors finite input separation even when the KL penalty is weak. Its retained information then decreases gently to a positive plateau: the strong-regularization limit stays input-dependent.

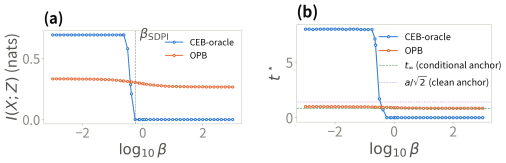

Figure 1: What strong regularization selects at noise rate $r = 0.2$: (a) $I(X;Z)$ against β for the CEB-oracle and OPB; (b) the optimal input-dependence amplitude $t^\star$ against β.

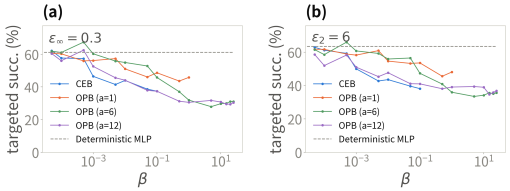

Figure 2: Targeted white-box PGD on MNIST (MLP): targeted success rate against β under (a) L_∞, $\varepsilon = 0.3$ and (b) L_2, $\varepsilon = 6$.

Table 2: Item-wise ablations on ImageNet-100 (MLP): for each contrast, the paired difference in accuracy points with its 95% bootstrap percentile CI and the variant's mean $\pm$ standard deviation.

factor	comparison	difference [95% CI]	variant mean $\pm$ std
readout	GPB $-$ GPB-H	$+0.96$ [0.77, 1.11]	83.44 ± 0.41
geometry	GPB $-$ GPB-G	$+1.26$ [0.96, 1.60]	83.14 ± 0.21
geometry convention	GPB $-$ GPB-CS	$+0.06$ [-0.03, 0.69]	84.35 ± 0.55
variance	GPB $-$ GPB-V	$+0.05$ [-0.27, 0.58]	84.35 ± 0.64
prior trainable	GPB $-$ GPB-P	$+1.02$ [0.45, 1.40]	83.38 ± 0.52
IB increment	GPB $-$ NCM-T	$+2.69$ [2.24, 3.00]	81.71 ± 0.50

Panel (b) reads the same selection geometrically, in the amplitude $t^\star$. The CEB-oracle's amplitude remains near the search bound at weak compression and drops to zero at the threshold, while the OPB amplitude converges to the conditional-expectation target t_∞ (0.8486 against $t_\infty = 0.8485$ at $\beta = 10^3$), never approaching the clean anchor $a/\sqrt{2}$: noisy labels pull the learned mean toward the expectation of the anchors. Strong regularization thus drives the posterior mean to the conditional-expectation anchor of Theorem 2.3(a) rather than to an input-independent representation.

3.4 Adversarial robustness

We evaluate robustness with the targeted MNIST PGD protocol of CEB (Fischer, 2020), always targeting class 0, with attack gradients and evaluation both on the stochastic prediction path (EOT-10 gradients, 10-sample MC predictions) rather than the posterior-mean inference used in the main comparison. Each curve is shown only where the stochastic clean accuracy exceeds 90% (Fig. 2).

Within that valid range, targeted success falls monotonically with compression: every curve starts at the deterministic baseline's level and declines by roughly a third to a half by the end of its valid range. The models differ in how far that range extends. CEB's curve ends at $\beta = 0.1$, where its success is 37.4%/38.1%; beyond it CEB's accuracy falls toward chance level, where near-zero success is trivial rather than a robustness gain, so no later point is plotted. OPB $a = 1$'s curve ends at $\beta = 1$, where its clean sampled-path accuracy drops below the 90% validity threshold. OPB $a = 6, 12$ stay valid over the whole grid and continue to fall, reaching 31.1%/30.5% under L_∞ and 35.7%/36.8% under L_2 at $\beta = 25$, roughly 6–7 points below CEB's last valid point under L_∞ and 1–2 under L_2, with clean accuracy still $\approx 98\%$; the curves plateau from $\beta \geq 5$ onward, so the gain sits in the compressed operating point rather than in β alone. Transfer attacks are reported as a partial masking check in Appendix A.4.3.

3.5 Item-wise ablations

The main comparison changes several factors simultaneously, so we also report item-wise ablations that approximately isolate six design factors, namely the prior geometry, the geometry convention, the readout, the prior variance, the trainability of the prior frame, and the IB term itself. Every variant is trained on ImageNet-100 (MLP backbone), and its results are reported at the best validation-selected configuration; all comparisons are paired by seed. Each row of Table 2 pairs two variants that differ in the factor named in its first column and reports the paired difference with its bootstrap CI together with the variant's mean; the variant definitions are in Appendix A.4.4.

Table 3: Posterior-geometry statistics: geometry concentration (mean absolute off-diagonal cosine of the class-center Gram; off-axis fraction for Housing), scale following (posterior center norm against the declared anchor scale; slope against ρ for Housing), tied metric (accuracy or R^2 under the tied head), and cross-seed stability (standard deviation of the concentration across runs).

Model	Geometry concentration	Scale following	Tied metric	Cross-seed stability
CEB (MNIST)	0.1205 ± 0.0026	5.14	98.16%	0.0201
OPB (MNIST)	0.0203 ± 0.0021	11.00	98.44%	0.0082
CEB (Housing)	0.0233	2.65 ± 0.59	$R^2 = 0.491$	0.593
EPB (Housing)	0.0002	9.22 ± 0.15	$R^2 = 0.778$	0.155

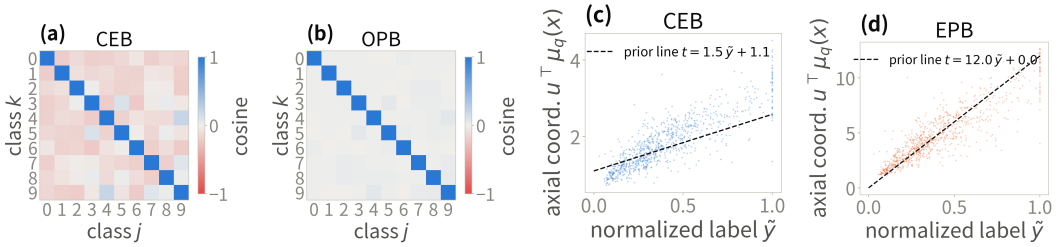

Figure 3: Posterior geometry: class-center Gram matrices for (a) CEB and (b) OPB, and axial coordinates for (c) CEB and (d) EPB.

The table shows the ablation results, in which two factors dominate. Removing the geometric construction alone, with the initialization, variance, and readout matched on both sides (GPB-G: free anchor means), costs 1.26 points, so the gain is the structural repair of the conditional bottleneck rather than an initialization or readout trick. GPB-CS, which keeps the trainable frame and centers the anchor table at every forward pass into a simplex ETF of the same pairwise spacing, is a null result, its confidence interval straddling zero: the coordinate convention is empirically immaterial, as the theory predicts, because the anchor sets differ only by a global translation that the readout and the conditional KL absorb once the posterior mean shifts, up to a $(1 - 1/K)$ rescaling of the anchor-norm-dependent constants. The gain therefore rests on the probabilistic coupling of the anchors to the KL target and the readout, not on the geometric primitive. The conditional compression is no passenger on that geometry: removing the stochastic bottleneck and its KL while keeping the frame and readout (NCM-T: deterministic output, no KL) still loses accuracy on its own. Freezing the prior frame (GPB-P) removes the co-adaptation on which the conditional KL relies, at real cost to accuracy. A free head (GPB-H) can refit the boundaries independently of the frame and undo part of the geometry, while the tied readout keeps prediction on it. The variance switch (GPB-V, learnable prior variance) is likewise a null result, so the fixed-covariance reference is empirically benign. Finer-grained decompositions of the ablation factors are reported in Appendix A.4.4.

3.6 Geometry transfer

Figure 3 reads the declared structure directly off the trained posteriors at matched compression $(\beta, a, \rho) = (0.1, 12, 12)$: OPB's class-center Gram matrix is visually an identity matrix while CEB's shows off-diagonal structure with no clean pattern, and on Housing EPB's axial scatter hugs its prior line while CEB's deviates and swings across runs. The posterior thus inherits the declared geometry; Table 3 reads the corresponding statistics, including the tied-metric payoff.

3.7 Compression sweeps

This subsection shows the classification sweep on ImageNet-100 (MLP): test accuracy and realized conditional KL against the extended β grid from $5 \cdot 10^{-5}$ to 25, for CEB and OPB at anchor scales $a \in \{1, 6, 12\}$; the California Housing regression companion is in Appendix A.4.5. The complementary scale-axis view at fixed β levels is reported in Appendix A.4.6.

Fig. 4(a) shows that larger anchor scales retain the high-accuracy plateau farther into strong compression: $a = 12$ keeps its accuracy where CEB has already left its valid high-accuracy range and

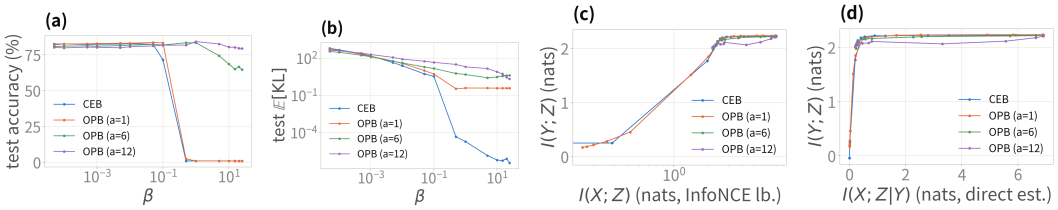

Figure 4: Compression behavior across two views and two tasks. (a)–(b) ImageNet-100 classification sweep (MLP): test accuracy and realized conditional KL against β. (c)–(d) MNIST information-plane and certificate-plane estimates.

$a = 1$ degrades toward chance accuracy, while at weak compression larger scales pay a small accuracy cost. The trade-off reverses under stronger compression, where the $a = 12$ scale is the most accurate configuration on the grid. By Corollary A.1 (Appendix A.2.2), this a=1 mean-path degradation is not demanded by the population target, which preserves the Bayes ordering at every β under isotropic posterior noise; it is a finite-capacity, optimization, or covariance effect. Fig. 4(b) shows where the two objectives spend the penalty: CEB's KL falls to near zero, reflecting the unconstrained-reference degeneracy, whereas OPB's KL stays anchored at the geometric scale of its fixed prior, with the retained plateau rising with the anchor scale. Together the curves make the anchor-scale/compression trade-off visible across the useful operating range.

3.8 Information-plane diagnostic

The MNIST information-plane estimates provide an auxiliary second view of the compression behavior. They are estimator-level diagnostics and do not carry the mechanism conclusions of this paper; the matched table and estimator details are in Appendix A.4.7, with the Housing companion there as well. The two planes show what each method discards as compression deepens. In the information plane (Fig. 4c, log $I(X;Z)$ axis), CEB and OPB with $a = 1$ follow the classic corner: task information is preserved at first, and past the corner $I(X;Z)$ and $I(Y;Z)$ fall together toward zero. OPB with $a = 6, 12$ likewise preserves task information early on, but past the corner compression appears to halt on a plateau just below the task-information ceiling $H(Y)$, with $I(Y;Z)$ flat. The certificate plane (Fig. 4d) shows the same contrast at the trajectories' endpoints: CEB and OPB with $a = 1$ reach zero residual only through the task-information collapse, whereas $a = 6, 12$ retain full task information on a stable nonzero residual plateau.

4 Conclusion and Limitations

Conclusion. We introduced the GPB, which declares the label-conditional geometry before training and ties the prediction head. The analysis proceeds in three steps: CEB's conditional KL is blind to the geometry it matches, and strong compression provably prefers an input-independent encoder (Theorem 2.1); the declared geometry instead makes posterior-center collapse strictly KL-expensive and excludes it above an explicit compression threshold (Theorem 2.2), though the contrast is at the objective level; and the modified objective drives the posterior mean to the conditional-expectation anchor, with the tied mean readouts Bayes-optimal at every compression strength and their excess risk bounded by anchor mismatch (Theorem 2.3; Appendix A.2.2). Across twelve dataset–backbone settings, GPB is best or tied-best in eight and never significantly worse than CEB, with the gain attributable primarily to the geometric prior and the IB term.

Limitations. Some limitations remain: the excess-risk bounds are population-level statements under an objective-comparison condition, not finite-sample bounds; the fixed orthogonal frame pins the code geometry rather than recovering semantic class similarities; and the adversarial study did not verify attack convergence. Making the geometry data-adaptive, so the anchors can reflect semantic class similarities without sacrificing the anti-collapse guarantees, is the most important next step. Proofs, supplementary evidence, and related work are collected in Appendices A.2–A.5.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish the writing, and used AI-based literature retrieval tools built on OpenAlex to assist the discovery of related work. All AI-assisted content was manually reviewed and verified by the authors; experimental results and theoretical claims were produced and checked by our own code and analysis. We take responsibility for the final content of this work, including text, claims or artifacts produced with generative AI.

REPRODUCIBILITY STATEMENT

The details needed to reproduce the results, including the twelve dataset–backbone settings, the data splits and preprocessing, the hyperparameter grids, and the evaluation and statistical protocols, are specified in Section 3.1, Appendix A.3.1, and Appendix A.4.1.

REFERENCES

Eslam Abdelaleem, Ilya Nemenman, and K. Michael Martini. Deep variational multivariate information bottleneck - a framework for variational losses. *Journal of Machine Learning Research*, 26(140):1–50, 2025. URL http://jmlr.org/papers/v26/24-0204.html.

Linara Adilova, Henning Petzka, Asja Fischer, and Bernhard C. Geiger. Geometric and information compression of representations in deep learning. In Michelangelo Ceci, Dragi Kocev, Anna Monreale, Celine Vens, Peter Flach, Giuseppe Manco, Myra Spiliopoulou, and Elio Masciari (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*, pp. 486–503, Cham, 2027. Springer Nature Switzerland. ISBN 978-3-032-37670-1.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck, 2019. URL https://arxiv.org/abs/1612.00410.

Bruno Belucci, Karim Lounici, and Katia Meziani. Adacap: An adaptive contrastive approach for small-data neural networks, 2025. URL https://arxiv.org/abs/2511.20170.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001/.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Weihong Deng, Jiani Hu, Xiuzhuang Zhou, and Jun Guo. Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning. *Pattern Recognition*, 47(12):3738–3749, 2014. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2014.06.020. URL https://www.sciencedirect.com/science/article/pii/S0031320314002428.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Toan Doan, Thin Nguyen, and Sunil Gupta. A geometric information bottleneck for activation steering. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 933–943, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3818109. URL https://doi.org/10.1145/3770855.3818109.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9), 2020. ISSN 1099-4300. doi: 10.3390/e22090999. URL https://www.mdpi.com/1099-4300/22/9/999.

Ian Fischer and Alexander A. Alemi. Ceb improves model robustness. *Entropy*, 22(10), 2020. ISSN 1099-4300. doi: 10.3390/e22101081. URL https://www.mdpi.com/1099-4300/22/10/1081.

Bernhard C. Geiger and Ian S. Fischer. A comparison of variational bounds for the information bottleneck functional. *Entropy*, 22(11), 2020. ISSN 1099-4300. doi: 10.3390/e22111229. URL https://www.mdpi.com/1099-4300/22/11/1229.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Huan Hua, Jun Yan, Xi Fang, Weiquan Huang, Huilin Yin, and Wancheng Ge. Causal information bottleneck boosts adversarial robustness of deep neural network, 2022. URL https://arxiv.org/abs/2210.14229.

John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, July 2012. ISSN 1549-9596. doi: 10.1021/ci3001277. URL https://doi.org/10.1021/ci3001277.

Yixin Ji, Jikai Wang, Juntao Li, Qiang Chen, Wenliang Chen, and Min Zhang. Early exit with disentangled representation and equiangular tight frame. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 14128–14142, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.889. URL https://aclanthology.org/2023.findings-acl.889/.

Ronald Katende. Geometry as a missing axis of representation quality: The variational geometric information bottleneck under data scarcity, 2026. URL https://arxiv.org/abs/2511.02496.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. ISSN 0167-7152. doi: https://doi.org/10.1016/S0167-7152(96)00140-X. URL https://www.sciencedirect.com/science/article/pii/S016771529600140X.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Artemy Kolchinsky, Brendan D. Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios, 2019a. URL https://arxiv.org/abs/1808.07593.

Artemy Kolchinsky, Brendan D. Tracey, and David H. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12), 2019b. ISSN 1099-4300. doi: 10.3390/e21121181. URL https://www.mdpi.com/1099-4300/21/12/1181.

Panagiotis Koromilas, Theodoros Giannakopoulos, Mihalis A. Nicolaou, and Yannis Panagakis. Neural collapse by design: Learning class prototypes on the hypersphere, 2026. URL https://arxiv.org/abs/2605.20302.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19538–19552. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a29a5ba2cb7bdeabba22de8c83321b46-Paper.pdf.

Tong Liang and Jim Davis. Inducing neural collapse to a fixed hierarchy-aware frame for reducing mistake severity. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1443–1452, 2023. doi: 10.1109/ICCV51070.2023.00139.

Lingkun Luo, Shiqiang Hu, and Liming Chen. Discriminative Noise Robust Sparse Orthogonal Label Regression-Based Domain Adaptation. *International Journal of Computer Vision*, 132(1): 161–184, January 2024. ISSN 1573-1405. doi: 10.1007/s11263-023-01865-z. URL https://doi.org/10.1007/s11263-023-01865-z.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1997–2005, 2017. doi: 10.1109/CVPRW.2017.250.

Gautam Pai, Alex Bronstein, Ronen Talmon, and Ron Kimmel. Deep isometric maps. *Image and Vision Computing*, 123:104461, 2022. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2022.104461. URL https://www.sciencedirect.com/science/article/pii/S0262885622000907.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2015509117.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008. doi: https://doi.org/10.1609/aimag.v29i3.2157. URL https://onlinelibrary.wiley.com/doi/abs/10.1609/aimag.v29i3.2157.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Siwei Wang and Stephanie E Palmer. Towards understanding neural collapse in supervised contrastive learning with the information bottleneck method, 2024. URL https://arxiv.org/abs/2305.11957.

Zifeng Wang, Tong Jian, Aria Masoomi, Stratis Ioannidis, and Jennifer Dy. Revisiting hilbert-schmidt information bottleneck for adversarial robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 586–597. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/055e31fa43e652cb4ab6c0ee845c8d36-Paper.pdf.

Cong Xu, Xiang Li, and Min Yang. An orthogonal classifier for improving the adversarial robustness of neural networks. *Information Sciences*, 591:251–262, 2022. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2022.01.039. URL https://www.sciencedirect.com/science/article/pii/S0020025522000627.

Hanqi Yan, Lin Gui, Menghan Wang, Kun Zhang, and Yulan He. Explainable recommender with geometric information bottleneck. *IEEE Transactions on Knowledge and Data Engineering*, 36 (7):3036–3046, 2024. doi: 10.1109/TKDE.2024.3350447.

Shujian Yu, Xi Yu, Sigurd Lø kse, Robert Jenssen, and Jose Principe. Cauchy-schwarz divergence information bottleneck for regression. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 40025–40068, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aecc44c30b1ea774c5d1d37da1a73b53-Paper-Conference.pdf.

Penglong Zhai and Shihua Zhang. Adversarial information bottleneck. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):221–230, 2024. doi: 10.1109/TNNLS.2022.3172986.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. Improving the adversarial robustness of NLP models by information bottleneck. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3588–3598, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.284. URL https://aclanthology.org/2022.findings-acl.284/.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

A APPENDIX

This appendix collects the notation, complete proofs, setup and construction details, additional experiments, and related work.

A.1 NOTATION

Each listed symbol has a single meaning throughout the paper; symbols used only inside a proof or an appendix subsection are defined locally at first use and are not listed.

Symbol	Object	Meaning
X, Y	variables	input and label random variables; deterministic-label settings write $Y = f(X)$
$Y - X - Z$	Markov chain	label–input–code Markov structure of the bottleneck
Z (z)	code	representation random variable and its realization
P_{XY}, P_Y	distributions	data distribution and label marginal
$\mathbb{E}$, $\mathbb{P}$, KL, CE, Var, diag, $\mathbf{1}\{\cdot\}$	operators	expectation, probability, KL divergence, cross-entropy, variance, diagonal operator, indicator
$H(\cdot)$, $I(\cdot;\cdot)$	information	entropy and mutual information, with subscript q when taken under the encoder-induced joint
$\mathcal{N}(\mu, \Sigma)$	Gaussian	multivariate Gaussian distribution
$q_\phi(z\|x)$	posterior	reparameterized Gaussian encoder distribution
$\mu_\phi(x)$, $\sigma^2_\phi(x)$	posterior mean / variance	learned diagonal posterior parameters
$p_\theta(z\|y)$	geometric prior	label-conditional Gaussian reference
$m_\theta(y)$	anchor mean	prior mean function (orthogonal anchors $a\, q_{\theta,k}$ for OPB, isometric axis $\rho\tilde{y}\, u$ for EPB)
τ^2	prior variance	fixed prior covariance scale, $\tau^2 = 1$
a, ρ	scales	anchor scale (OPB) / isometric scale (EPB)
C_{col}	collapse cost	$\frac{a^2}{2\tau^2}\left(1 - \sum_k \pi_k^2\right)$ (OPB) / $\frac{\rho^2}{2\tau^2}\,\mathrm{Var}(\tilde{Y})$ (EPB)
U_θ, Q_θ, $q_{\theta,k}$	matrices	all-class raw matrix, its polar frame, and the k-th frame column (OPB)
u	unit direction	EPB normalized trainable direction
W	direction matrix	EPB trainable direction before normalization
$s_k(z)$	score	anchor-energy score of class k at representation z
$\tilde{y}$ $(\tilde{Y})$	normalized label	min–max normalized target (and its random variable)
$\hat{\tilde{y}}$	prediction	tied projection of the regression head
π_k	class priors	label probabilities
K, d	dimensions	number of classes / latent dimension
$\mathbb{R}^d$, $\mathcal{Y}$	spaces	d-dimensional latent space and label space
d_Y	label metric	metric on the label space
β	KL weight	compression weight in the objective
ℓ, c_ψ	loss, readout	prediction loss and classifier/regression head
w, $\widehat{w}$, w_0	models	trainable parameters, trained solution, comparator
$\mathcal{L}_{\mathrm{GPB}}$	objective	GPB training objective, prediction risk plus β times the expected KL
$\mathcal{L}_{\mathrm{CEB}}$ $(\mathcal{L}^\star_{\mathrm{CEB}})$	objective	CEB objective and its idealized nested form $\min_{r,c}[\mathcal{R}(q,c) + \beta\mathcal{K}(q,r)]$
$\mathcal{R}(w)$, $\mathcal{K}(w)$	risk, KL	prediction risk and expected KL of model w
$\mathcal{R}(q,c)$, $\mathcal{K}(q,r)$	risk, KL	risk of readout c and expected KL of encoder q against reference r
R_0	comparator risk	finite risk of the conditional-anchor comparator
$q(z\|y)$	aggregated posterior	label-conditional mixture of the encoder, $\int q(z\|x)p(x\|y)\,dx$; $q(y\|z)$ the induced label posterior
Δ, $\bar{\Delta}$	declared separation	minimal/maximal anchor distance of the discrete-label separation condition
$\rho_{\min}$, $\rho_{\max}$	Lipschitz constants	bi-Lipschitz bounds of the anchor map
$\mathcal{G}$	family	geometric prior family of the GPB framework
$\eta_k(x)$ $(\eta(x))$	label probability	$\mathbb{P}(Y = k \mid X = x)$ and its vector
$\nu(X)$	conditional mean	$\mathbb{E}[\tilde{Y} \mid X]$ (EPB)
$\bar{m}(x)$	conditional anchor	$\mathbb{E}[m_\theta(Y) \mid X = x]$, the conditional-expectation anchor

Symbol	Object	Meaning
δ_m, $A_m(w)$	mismatch parts	label-ambiguity and mean-approximation parts of the KL decomposition; δ_m equals $\frac{a^2}{2\tau^2}\mathbb{E}[1 - \|\eta(X)\|_2^2]$ for OPB and $\frac{\rho^2}{2\tau^2}\mathbb{E}[\mathrm{Var}(\tilde{Y} \mid X)]$ for EPB
$V(\Sigma)$	variance term	expected variance-mismatch term of the conditional KL, $\frac{1}{2}\mathbb{E}[\mathrm{tr}(\Sigma/\tau^2) - d - \log\det(\Sigma/\tau^2)]$
$\mu^\star_\beta(x)$	optimal mean	population-optimal posterior mean at compression weight β
ε	tolerance	optimality tolerance of the training objective
t, $t^\star$	amplitude	input-dependence amplitude of the controlled encoder family and its optimum
t_∞	target	conditional-expectation target $a(1-2r)/\sqrt{2}$ of the controlled experiment
β_{SDPI}	threshold	strong data-processing threshold $(1-2r)^2/(4r(1-r))$ of the controlled experiment

A.2 Proofs and theoretical analysis

A.2.1 Degeneracy analysis and anti-collapse guarantees

Proof of Theorem 2.1(a). Throughout, "collapse" refers only to coincident class-conditional means, not to independence $Z \perp Y$ or $I(Y;Z) = 0$; the translated pair of part (a) is $q'(z|x) = q(z - t_{f(x)} \,|\, x)$ and $r'(z|y) = r(z - t_y \,|\, y)$. Write T_y for the translation $z \mapsto z + t_y$ and $T_{\#}P$ for the pushforward of a distribution P under a map T. The density of $T_y\#P$ at z is $p(z - t_y)$, and the substitution $w = z - t_y$ in the KL integral gives

$$\mathrm{KL}(T_{\#}P \,\|\, T_{\#}Q) = \int p(z - t_y) \log \frac{p(z - t_y)}{q(z - t_y)}\, dz = \int p(w) \log \frac{p(w)}{q(w)}\, dw = \mathrm{KL}(P \,\|\, Q),$$

so KL is invariant under a common translation of its two arguments. Since $Y = f(X)$ is deterministic, each sample (x, y) carries the single label $y = f(x)$, and both arguments of its per-sample KL term receive the same map T_y:

$$\mathrm{KL}(q'(z|x) \,\|\, r'(z|y)) = \mathrm{KL}(T_{y\#}q(z|x) \,\|\, T_{y\#}r(z|y)) = \mathrm{KL}(q(z|x) \,\|\, r(z|y)).$$

Averaging over $p(x, y)$ gives

$$\mathcal{K}(q', r') = \mathbb{E}_{p(x,y)}\, \mathrm{KL}(q'(z|x) \,\|\, r'(z|y)) = \mathbb{E}_{p(x,y)}\, \mathrm{KL}(q(z|x) \,\|\, r(z|y)) = \mathcal{K}(q, r)$$

for every choice of $\{t_y\}$: the certificate is invariant under per-label translations.

It remains to check that the translations realize both geometries: the transformed class-conditional mean is $\bar{m}_y + t_y$, where $\bar{m}_y = \mathbb{E}_{q(z|y)}[Z]$. Collapse: $t_y = c_0 - \bar{m}_y$ puts every class mean at c_0. Separation: take $t_y = a\, v_y$ for an orthonormal set $\{v_y\}$ (available since $d \geq K$ in all settings of this paper). The transformed class-conditional means are

$$m'_y = \bar{m}_y + t_y = \bar{m}_y + a\, v_y,$$

each class mean pushed by a along its own orthogonal direction. The difference of two transformed means is

$$m'_y - m'_{y'} = (\bar{m}_y - \bar{m}_{y'}) + (t_y - t_{y'}) = (\bar{m}_y - \bar{m}_{y'}) + a(v_y - v_{y'}),$$

the fixed original gap plus the imposed shift difference, whose length is $\|v_y - v_{y'}\|\, a = \sqrt{2}\, a$ by orthonormality. The reverse triangle inequality, applied to these two summands, then gives

$$\|m'_y - m'_{y'}\| \geq \|t_y - t_{y'}\| - \|\bar{m}_y - \bar{m}_{y'}\| = \sqrt{2}\, a - \|\bar{m}_y - \bar{m}_{y'}\|,$$

unbounded in a: the shift difference dominates the fixed gap, so any prescribed separation is exceeded for a large enough. Both choices leave $\mathcal{K}$ unchanged, so, since the starting pair was arbitrary, every attainable certificate value admits both a collapsed and an arbitrarily separated realization of the class-conditional means: the certificate value alone cannot determine which geometry was learned. Feasibility in the diagonal-Gaussian families used here is immediate: translations shift the means and preserve the diagonal covariance, so $r'(z|y)$ remains in the reference family and $q'(z|x)$ remains a Gaussian encoder, the shift $t_{f(x)}$ being well defined because $Y = f(X)$ is deterministic. □

Proof of Theorem 2.1(b). With sufficiently expressive families, the joint minimization over (r, c) separates, since the risk involves only the readout and the KL only the reference:

$$\mathcal{L}^{\star}_{\mathrm{CEB}}(q) = \min_{c} \mathcal{R}(q, c) + \beta \min_{r} \mathcal{K}(q, r).$$

The first inner minimization is solved by the Bayes readout $c(y|z) = q(y|z)$, where $q(y|z)$ is the label posterior induced by the encoder $q(z|x)$, the optimal log-loss predictor, with value $H_q(Y|Z)$; the second by the aggregated reference $r(z|y) = q(z|y)$ (Fischer, 2020), with value $I_q(X; Z|Y)$. Hence

$$\mathcal{L}^{\star}_{\mathrm{CEB}}(q) = H_q(Y|Z) + \beta I_q(X; Z|Y), \tag{18}$$

and, since $Y - X - Z$ gives $I(Y; Z|X) = 0$, the chain rule yields $I(X; Z|Y) = I(X; Z) - I(Y; Z)$, so

$$\mathcal{L}^{\star}_{\mathrm{CEB}}(q) = H(Y) + \beta I(X; Z) - (1 + \beta) I(Y; Z). \tag{19}$$

By definition of $\bar{\eta}$ as a supremum, every processing U of X with $I(X; U) > 0$ satisfies the coefficient bound $I(Y; U) \leq \bar{\eta}\, I(X; U)$. The encoder maps X to Z without access to Y, so $Z - X - Y$ holds and the bound applies with $U = Z$:

$$I(Y; Z) \leq \bar{\eta}\, I(X; Z).$$

Since the coefficient of $I(Y; Z)$ in Eq. (19) is negative, this inequality flips direction upon substitution:

$$\mathcal{L}^{\star}_{\mathrm{CEB}}(q) \geq H(Y) + \beta I(X; Z) - (1 + \beta)\bar{\eta}\, I(X; Z) = H(Y) + \big[\beta - (1 + \beta)\bar{\eta}\big] I(X; Z).$$

The bracket is non-negative exactly when $\beta \geq \bar{\eta}/(1 - \bar{\eta})$ (the algebra uses $\bar{\eta} < 1$), and $I(X; Z) \geq 0$, hence

$$\mathcal{L}^{\star}_{\mathrm{CEB}}(q) \geq H(Y).$$

For the input-independent posterior $q_{\mathrm{col}}(z|x) = p_0(z)$, the joint factorizes as $p(z, y) = p_0(z)\, p(y)$, so Z is independent of (X, Y) and $I(X; Z) = 0$ and $I(Y; Z) = 0$, giving

$$\mathcal{L}^{\star}_{\mathrm{CEB}}(q_{\mathrm{col}}) = H(Y),$$

attained by the Bayes readout and the aggregated reference. Thus $\mathcal{L}^{\star}_{\mathrm{CEB}}(q) \geq \mathcal{L}^{\star}_{\mathrm{CEB}}(q_{\mathrm{col}})$ for every posterior q, and the input-independent posterior is a global minimizer. At strict inequality $\beta > \bar{\eta}/(1 - \bar{\eta})$ the bracket in Eq. (19) is strictly positive, so every posterior with $I(X; Z) > 0$ satisfies $\mathcal{L}^{\star}_{\mathrm{CEB}}(q) > H(Y)$ and is not a minimizer: every global minimizer satisfies $I(X; Z) = 0$. At equality, the bracket vanishes and the input-independent posterior is one global minimizer. □

The condition requires label ambiguity, captured by $\bar{\eta} < 1$; for deterministic labels an expressive encoder attains $I(Y; Z) = I(X; Z)$ (take $Z = f(X)$), so $\bar{\eta} = 1$ and no finite β triggers the threshold, consistent with Theorem 2.1(a). For the binary symmetric label channel $Y = X \oplus N$ with $N \sim \mathrm{Bernoulli}(r_{\mathrm{n}})$, the strong data-processing coefficient is the classical value $\bar{\eta} = (1 - 2r_{\mathrm{n}})^2$, and the threshold reads

$$\beta \geq \frac{(1 - 2r_{\mathrm{n}})^2}{4r_{\mathrm{n}}(1 - r_{\mathrm{n}})}, \tag{20}$$

which decreases from ∞ to 0 as r_{n} grows from 0 to $1/2$, for example $\beta \geq 1.78$ at $r_{\mathrm{n}} = 0.1$, $\beta \geq 0.56$ at $r_{\mathrm{n}} = 0.2$, and $\beta \geq 0.042$ at $r_{\mathrm{n}} = 0.4$. Part (b) itself is channel-general: any label channel with $\bar{\eta} < 1$ triggers the threshold. The mutual-information coefficient $\bar{\eta}$ must not be confused with the square of the channel's Dobrushin coefficient: the two coincide for binary symmetric channels but not for K-ary ones, where the Dobrushin square understates $\bar{\eta}$. At $K = 10$, $r_{\mathrm{n}} = 0.4$, and uniform X, the input-preserving choice $U = X$ in Eq. (3) attains $I(Y; U)/I(X; U) = [\log K - h_2(r_{\mathrm{n}}) - r_{\mathrm{n}} \log(K - 1)]/\log K \approx 0.327$, which already exceeds $(1 - r_{\mathrm{n}}K/(K - 1))^2 \approx 0.309$. The controlled experiment of Sec. 3.3 therefore reports the qualitative direction of the threshold (it decreases as the noise grows) rather than K-ary closed-form values; there the ambiguity cost is read from the GPB side (the OPB/EPB forms of δ_m in Appendix A.2.2), while Eq. (20) describes the binary unconstrained-reference side.

Scope of the idealized analysis. These statements concern idealized objectives in which the reference and readout are optimized within sufficiently expressive families and the optima are only approximately attained by joint non-convex training. The implemented CEB instantiates the setup with a learned per-label Gaussian reference (a linear prior network with per-label mean and variance) and a learned readout; the per-label translations of Theorem 2.1(a) act on the reference side exactly, and $\mathcal{N}(0, I)$ lies in the reference family whenever the prior network maps all labels to a common mean and variance, so the zero-KL endpoint of the certificate ambiguity is attainable in the implementation. Accordingly, part (a) is an invariance statement about the certificate: it does not certify that a particular run reaches collapse, and on deterministic benchmarks separated configurations coexist with collapsed ones at every certificate value, so the collapse observed at large β (Secs. 3.7 and 3.8) is an optimization- and parameterization-dependent selection among certificate-equivalent geometries, not an unconditional consequence of the objective. Likewise, the threshold of Theorem 2.1(b) is a sufficient condition for the idealized objective, not a predicted exact phase transition of finite training runs. $H(Y|X) > 0$ alone does not imply $\bar{\eta} < 1$—with $X = A$ and $Y = (A, N)$ for independent noise N, the representation $U = A$ already achieves $I(Y;U) = I(X;U)$—so part (b) applies to label channels satisfying the strict strong data-processing condition, and neither noisy labels in general nor a finite training set in which every input appears with a single label establishes $\bar{\eta} < 1$.

For a reference family $\mathcal{P}$, the infimum certificate $\mathcal{K}_{\mathcal{P}}(q) = \inf_{p\in\mathcal{P}} \mathbb{E}_{p(x,y)} \mathrm{KL}(q(z \mid x) \,\|\, p(z \mid y))$ satisfies $\mathcal{K}_{\mathcal{P}}(q) = I_q(X; Z \mid Y) + \inf_{p\in\mathcal{P}} \mathbb{E}_{p(y)} \mathrm{KL}(q(z \mid y) \,\|\, p(z \mid y))$, because Eq. (2) holds for every fixed reference and its first summand is independent of p; the infimum of the non-negative second term can only increase as the family shrinks, so restricting the reference family never tightens the certificate for a fixed encoder.

Proof of Theorem 2.2(a). If all class-conditional centers lie at a common point c_0, that is $\bar{\mu}(Y) = \mathbb{E}[\mu_\phi(X) \mid Y] = c_0$, the conditional variance decomposition gives

$$\mathbb{E}\|\mu_\phi(X) - A_Y\|^2 = \mathbb{E}\|\mu_\phi(X) - c_0\|^2 + \mathbb{E}\|c_0 - A_Y\|^2,$$

the cross term vanishing since $\mathbb{E}[\mu_\phi(X) - c_0 \mid Y] = 0$. The standard diagonal-Gaussian KL formula splits the KL into its mean and variance terms,

$$\mathcal{K}(q, r) = \frac{1}{2\tau^2}\mathbb{E}\|\mu_\phi(X) - A_Y\|^2 + V_0,$$

where $A_Y := aq_{\theta,Y}$ is the random anchor of the sampled label, and $V_0 \geq 0$ is the non-negative variance term of that formula. Hence

$$\mathcal{K}(q, r) \geq \frac{1}{2\tau^2} \min_{c_0} \mathbb{E}\|c_0 - A_Y\|^2.$$

The mean term is minimized by completing the square. Writing $c - A_Y = (c - \mathbb{E}A_Y) - (A_Y - \mathbb{E}A_Y)$ and expanding the squared norm, the cross term vanishes because $\mathbb{E}[A_Y - \mathbb{E}A_Y] = 0$:

$$\mathbb{E}\|c - A_Y\|^2 = \|c - \mathbb{E}A_Y\|^2 + \mathbb{E}\|A_Y - \mathbb{E}A_Y\|^2.$$

The second term is the variance of A_Y by definition:

$$\mathbb{E}\|A_Y - \mathbb{E}A_Y\|^2 = \mathrm{Var}(A_Y),$$

so the minimum over c_0 is $\mathrm{Var}(A_Y)$, attained at the anchor centroid $c_0^\star = \mathbb{E}A_Y = a\bar{q}$ with $\bar{q} = \sum_k \pi_k q_{\theta,k}$. The anchors are unit vectors, so $\mathbb{E}\|A_Y\|^2 = a^2$, and orthonormality gives $\|\mathbb{E}A_Y\|^2 = a^2\|\bar{q}\|^2 = a^2 \sum_k \pi_k^2$, hence

$$\mathrm{Var}(A_Y) = \mathbb{E}\|A_Y\|^2 - \|\mathbb{E}A_Y\|^2 = a^2\Big(1 - \sum_k \pi_k^2\Big),$$

which is strictly positive since $\sum_k \pi_k = 1$ and at least two labels have positive probability, proving

$$\mathcal{K}(q, r) \geq \frac{a^2}{2\tau^2}\Big(1 - \sum_{k=1}^{K} \pi_k^2\Big) > 0, \tag{21}$$

The input-independent posterior $\mathcal{N}(c_0^\star, \tau^2 I)$ attains the bound, its variance term vanishing. For EPB the analogous statement holds with axis anchors $A_Y = \rho\tilde{Y}u$: the minimizer is $c_0 = \rho\mathbb{E}[\tilde{Y}]u$ and the cost $\frac{\rho^2}{2\tau^2}\mathrm{Var}(\tilde{Y}) > 0$ whenever labels vary. □

Throughout this appendix, fitted models are compared against a prespecified feasible comparator w_0 through the *objective-comparison condition* $\mathcal{L}_\beta(\widehat{w}) \le \mathcal{L}_\beta(w_0) + \varepsilon$ on the population target $\mathcal{L}_\beta = \mathcal{R} + \beta\mathcal{K}$; global ε-optimality is a sufficient instance.

Proof of Theorem 2.2(b). We prove the fitted-model version of part (b): if w_0 satisfies $\mathcal{K}(w_0) < C_{\mathrm{col}}$ and the fitted model $\widehat{w}$ satisfies the objective-comparison condition, then $\widehat{w}$ avoids complete center coincidence whenever $\beta > \frac{\mathcal{R}(w_0)+\varepsilon}{C_{\mathrm{col}}-\mathcal{K}(w_0)}$; with $\varepsilon = 0$, no center-collapsed model is globally optimal. *Collapse floor.* For any center-collapsed w, nonnegativity of $\mathcal{R}$ and part (a) give

$$\mathcal{L}_\beta(w) = \mathcal{R}(w) + \beta\mathcal{K}(w) \ge \beta C_{\mathrm{col}} :$$

center collapse prices every collapsed model at least βC_{col}.

Threshold reformulation. Since $\mathcal{K}(w_0) < C_{\mathrm{col}}$, the denominator $C_{\mathrm{col}} - \mathcal{K}(w_0)$ is positive and the condition $\beta > \frac{\mathcal{R}(w_0)+\varepsilon}{C_{\mathrm{col}}-\mathcal{K}(w_0)}$ is equivalent to

$$\beta C_{\mathrm{col}} > \mathcal{R}(w_0) + \beta\mathcal{K}(w_0) + \varepsilon = \mathcal{L}_\beta(w_0) + \varepsilon :$$

the collapse floor exceeds the comparator objective up to tolerance.

Chaining. Combining the objective-comparison condition $\mathcal{L}_\beta(\widehat{w}) \le \mathcal{L}_\beta(w_0) + \varepsilon$ with the two displays gives

$$\mathcal{L}_\beta(\widehat{w}) \le \mathcal{L}_\beta(w_0) + \varepsilon < \beta C_{\mathrm{col}} \le \mathcal{L}_\beta(w)$$

for every center-collapsed w, so $\widehat{w}$ lies strictly below every collapsed model and can itself not be collapsed.

Global optimality ($\varepsilon = 0$). With $\varepsilon = 0$ the threshold reformulation gives $\beta C_{\mathrm{col}} > \mathcal{L}_\beta(w_0)$, so every center-collapsed w satisfies $\mathcal{L}_\beta(w) \ge \beta C_{\mathrm{col}} > \mathcal{L}_\beta(w_0)$ and is strictly dominated by the feasible comparator w_0: no center-collapsed model is globally optimal. □

Predictable-structure margin. For the conditional-anchor comparator $w_0 = (\bar{m}, \tau^2 I)$ of Theorem 2.3(a), the margin is $C_{\mathrm{col}} - \mathcal{K}(w_0) = \frac{a^2}{2\tau^2}\mathbb{E}\|\eta(X) - \pi\|_2^2 =: G$ (zero approximation and covariance error in the decomposition of Eq. (13), using the OPB form of δ_m of Appendix A.2.2), positive exactly when the input changes the label posterior on a positive-probability set. With mean-approximation and covariance errors totaling A_0, the margin is $G - A_0$ and the exclusion condition reads $G > A_0$ and $\beta(G - A_0) > \mathcal{R}(w_0) + \varepsilon$: the predictable label structure must outweigh the model's approximation and mismatch costs. For EPB the same comparator, with $\bar{m} = \rho\,\mathbb{E}[\tilde{Y}|X = x]\,u$, gives $C_{\mathrm{col}} - \mathcal{K}(w_0) = \frac{\rho^2}{2\tau^2}\,\mathrm{Var}(\mathbb{E}[\tilde{Y}|X])$; the guarantee concerns the predictable mean structure and does not cover label dependence carried only by the conditional variance.

A.2.2 Proofs and quantitative refinements for Theorem 2.3

Proof of part (a): strong-regularization limit. The comparator w_0 with posterior $\mathcal{N}(\bar{m}(x), \tau^2 I)$ attains $\mathcal{K}(w_0) = \delta_m$ by the decomposition of Eq. (13) ($A_m = 0$, $V = 0$). Since the risk is nonnegative, the objective-comparison condition gives

$$\beta\big(\delta_m + A_m(\widehat{w}) + V(\hat{\Sigma})\big) \le \beta\mathcal{K}(\widehat{w}) \le \mathcal{R}(\widehat{w}) + \beta\mathcal{K}(\widehat{w}) \le R_0 + \beta\delta_m + \varepsilon,$$

so $A_m(\widehat{w}) + V(\hat{\Sigma}) \le (R_0 + \varepsilon)/\beta$, which is Eq. (15). For OPB and EPB the ambiguity constant reads $\delta_m = \frac{a^2}{2\tau^2}\mathbb{E}[1 - \|\eta(X)\|_2^2]$ and $\delta_m = \frac{\rho^2}{2\tau^2}\mathbb{E}[\mathrm{Var}(\tilde{Y} \mid X)]$, from the anchor-variance identity $\mathbb{E}[\|aq_{\theta,Y} - aQ_\theta\eta(X)\|_2^2 \mid X] = a^2(1 - \|\eta(X)\|_2^2)$ and its EPB analogue $\rho^2\,\mathrm{Var}(\tilde{Y} \mid X)$.

Proof of part (b), OPB: characterization. Write $\xi \sim \mathcal{N}(0, \Sigma_\theta(x))$ for the posterior noise and $p_\mu(\xi; x) = \mathrm{softmax}\big(\frac{a}{\tau^2}Q_\theta^\top(\mu + \xi)\big)$, and

$$F_x(\mu) = \mathbb{E}_\xi\,\mathrm{CE}\big(\eta(x),\, p_\mu(\xi; x)\big) + \frac{\beta}{2\tau^2}\|\mu - aQ_\theta\eta(x)\|_2^2, \tag{22}$$

the mean-dependent part of the population objective conditional on x at fixed posterior covariance $\Sigma_\theta(x)$ (the dropped terms of Eq. (13) are μ-independent; the noise is drawn conditionally on $X = x$

alone, so $\mathbb{E}[\xi \mid X, Y] = 0$). For fixed ξ the logits are affine in μ, and $\mathrm{CE}(\eta, \mathrm{softmax}(\cdot)) = -\eta^\top(\cdot) + \log \sum_k e^{(\cdot)_k}$ is convex in its argument (a linear term plus the log-sum-exp), so the risk term is convex in μ; the anchor term has Hessian $(\beta/\tau^2)I$, so F_x is (β/τ^2)-strongly convex, hence coercive, and admits a unique minimizer $\mu^\star_\beta(x)$. For fixed ξ, $\nabla_\mu \log p_{\mu,k}(\xi; x) = \frac{a}{\tau^2} Q_\theta(e_k - p_\mu(\xi; x))$, so

$$\nabla_\mu \mathrm{CE}\big(\eta(x), p_\mu(\xi; x)\big) = -\frac{a}{\tau^2} Q_\theta \sum_k \eta_k(x)\big(e_k - p_\mu(\xi; x)\big) = \frac{a}{\tau^2} Q_\theta\big(p_\mu(\xi; x) - \eta(x)\big),$$

using $\sum_k \eta_k(x) = 1$; the gradient is uniformly bounded ($\|p_\mu(\xi; x) - \eta(x)\|_2 \le \sqrt{2}$), so dominated convergence licenses exchanging gradient and expectation. The first-order condition $0 = \nabla F_x(\mu^\star_\beta(x))$ therefore reads

$$0 = \frac{a}{\tau^2} Q_\theta\big(\bar{p}_\beta(x) - \eta(x)\big) + \frac{\beta}{\tau^2}\big(\mu^\star_\beta(x) - a Q_\theta \eta(x)\big),$$

and rearranging gives the fixed point

$$\mu^\star_\beta(x) = a Q_\theta \eta(x) + \frac{a}{\beta} Q_\theta\big(\eta(x) - \bar{p}_\beta(x)\big), \qquad \bar{p}_\beta(x) = \mathbb{E}_\xi p_{\mu^\star_\beta}(\xi; x), \tag{23}$$

the geometric embedding of the label posterior plus an $O(\beta^{-1})$ correction; constant score priors cancel in its gradient, and the closed form is a linear image of Q_θ, so $\mu^\star_\beta(x) \in \mathrm{col}(Q_\theta)$. Since $\eta(x)$ and $\bar{p}_\beta(x)$ are probability vectors and Q_θ is an isometry on its columns, the fixed point gives Eq. (16), so $A_m(w^\star_\beta) \le a^2/(\tau^2\beta^2)$ and, by Eq. (27), $\mathcal{E}_{\mathrm{cls}}(g_{\mu^\star_\beta}) \le \min\{1,\ 2/\beta\}$ (the excess risk $\mathcal{E}_{\mathrm{cls}}$ and the mean readout are defined in the predictive-consequence remark below): the population risk of the optimal mean-path readout decays as $O(\beta^{-1})$. The fixed point describes the population target, not the trajectory of non-convex training.

Corollary A.1 (Order preservation under isotropic posterior noise)**.** *Suppose $\Sigma_\theta(x) = s^2(x)I$. With $v(x) = Q_\theta^\top \mu(x)/a$, Eq. (23) reads*

$$\beta v^\star + \bar{p}_\beta(v^\star) = (\beta + 1)\eta(x), \tag{24}$$

where $\bar{p}_\beta(v) = \mathbb{E}_{\tilde{\xi}} \mathrm{softmax}\big(\frac{a^2}{\tau^2} v + \tilde{\xi}\big)$ and, under the isotropic covariance, $\tilde{\xi} = \frac{a}{\tau^2} Q_\theta^\top \xi$ has i.i.d. coordinates (the logits of Eq. (22) being $\frac{a}{\tau^2} Q_\theta^\top(\mu + \xi) = \frac{a^2}{\tau^2} v + \tilde{\xi}$). Then for every pair $i \ne j$,

$$\mathrm{sign}(v^\star_i - v^\star_j) = \mathrm{sign}(\eta_i(x) - \eta_j(x)), \tag{25}$$

and hence, for the energy scores of Eq. (10) with uniform priors $\pi_k = 1/K$, $\arg\max_k q_{\theta,k}^\top \mu^\star_\beta(x) = \arg\max_k \eta_k(x)$ for every $\beta > 0$: at fixed isotropic posterior noise, uniform score priors, and a fully expressive mean function, the optimal mean-path readout is Bayes-optimal at every compression strength.

Proof. Multiplying Eq. (23) by $Q_\theta^\top/a$ and writing $v^\star(x) = Q_\theta^\top \mu^\star_\beta(x)/a$ gives Eq. (24), where under $\Sigma_\theta(x) = s^2(x)I$ the projected noise $\tilde{\xi} = \frac{a}{\tau^2} Q_\theta^\top \xi$ has covariance $\frac{a^2 s^2(x)}{\tau^4} Q_\theta^\top Q_\theta = \frac{a^2 s^2(x)}{\tau^4} I_K$, so its coordinates are i.i.d. We first record that the expected softmax preserves the order of its logits under exchangeable noise: for $i \ne j$, fix all coordinates except i, j and parameterize v by $\delta = v_i - v_j$ with $v_i + v_j$ held fixed. Write $g(\tilde{\xi}, \delta) = \mathrm{softmax}_i(v + \tilde{\xi}) - \mathrm{softmax}_j(v + \tilde{\xi})$ and $h(\delta) = \mathbb{E} g(\tilde{\xi}, \delta)$. With $p = \mathrm{softmax}(v + \tilde{\xi})$,

$$\frac{\partial g}{\partial \delta} = \frac{1}{2}\Big(\frac{\partial g}{\partial v_i} - \frac{\partial g}{\partial v_j}\Big) = \frac{1}{2}\big(p_i(1 - p_i + p_j) + p_j(1 - p_j + p_i)\big) = \frac{1}{2}\big(p_i(1 - p_i) + p_j(1 - p_j) + 2p_i p_j\big) > 0,$$

since $0 < p_i, p_j < 1$; bounded convergence gives $h'(\delta) = \mathbb{E}[\partial g/\partial \delta] > 0$, so h is strictly increasing. At $\delta = 0$ the two coordinates of v coincide, and swapping $\tilde{\xi}_i$ with $\tilde{\xi}_j$ preserves the noise law while exchanging the two softmax entries, so $\bar{p}_i(v) = \bar{p}_j(v)$ and $h(0) = 0$: the expected softmax preserves the order of its logits. For a pair $i \ne j$, subtracting the j-th coordinate of Eq. (24) from the i-th gives

$$\beta(v^\star_i - v^\star_j) + \big(\bar{p}_{\beta,i}(v^\star) - \bar{p}_{\beta,j}(v^\star)\big) = (\beta + 1)\big(\eta_i(x) - \eta_j(x)\big),$$

and the order-preservation argument applied to the logits $\frac{a^2}{\tau^2} v^\star$ (a positive rescaling of $v^\star$) makes both terms on the left share the sign of $v^\star_i - v^\star_j$; hence $\mathrm{sign}(v^\star_i - v^\star_j) = \mathrm{sign}(\eta_i(x) - \eta_j(x))$ for every pair, which is Eq. (25). □

Proof of part (b), EPB: exact solution under squared loss. Fix the unit direction u, let $f(x) = u^\top \mu(x)/\rho$ and $\mu_\perp(x) = (I - uu^\top)\mu(x)$, and let $V(\Sigma)$ be the variance-mismatch term of the decomposition of Eq. (13). The population EPB objective decomposes exactly as

$$\mathcal{L}_\beta = \Big(1 + \frac{\beta\rho^2}{2\tau^2}\Big)\mathbb{E}\big(f(X) - \tilde{Y}\big)^2 + \frac{\beta}{2\tau^2}\mathbb{E}\|\mu_\perp(X)\|_2^2 + \frac{1}{\rho^2}\mathbb{E}\big[u^\top \Sigma_\theta(X) u\big] + \beta V(\Sigma_\theta).$$

Exact decomposition. With $Z = \mu(X) + \xi$ and the noise ξ drawn from $\mathcal{N}(0, \Sigma_\theta(X))$ conditionally on X alone, so that $\mathbb{E}[\xi \mid X, Y] = 0$, the squared loss expands as

$$\mathbb{E}\Big[\Big(\frac{u^\top Z}{\rho} - \tilde{Y}\Big)^2\Big] = \mathbb{E}\big(f(X) - \tilde{Y}\big)^2 + \frac{1}{\rho^2}\mathbb{E}\big[u^\top \Sigma_\theta(X) u\big],$$

the cross term vanishing because $\mathbb{E}\big[(f(X) - \tilde{Y})u^\top \xi\big] = \mathbb{E}\big[(f(X) - \tilde{Y})u^\top \mathbb{E}[\xi \mid X, Y]\big] = 0$. The KL mean term decomposes along u and its complement:

$$\mathbb{E}\|\mu(X) - \rho\tilde{Y}u\|_2^2 = \mathbb{E}\|\mu_\perp(X)\|_2^2 + \rho^2\mathbb{E}\big(f(X) - \tilde{Y}\big)^2,$$

and the variance part is $V(\Sigma_\theta)$. Summing the risk with β times the KL proves the decomposition.

Optimal mean. The map $\mu \mapsto (f, \mu_\perp)$ with $f = u^\top\mu/\rho$ and $\mu_\perp = (I - uu^\top)\mu$ is a bijection (inverse $\mu = \rho f u + \mu_\perp$), so the mean terms are minimized pointwise: with $\nu(X) = \mathbb{E}[\tilde{Y} \mid X]$,

$$\mathbb{E}\big(f(X) - \tilde{Y}\big)^2 = \mathbb{E}\big(f(X) - \nu(X)\big)^2 + \mathbb{E}\,\mathrm{Var}(\tilde{Y} \mid X),$$

so the f-term is minimized at $f^\star(x) = \nu(x)$ (its coefficient $1 + \beta\rho^2/(2\tau^2) > 0$), while the $\mu_\perp$-term is non-negative and vanishes exactly at $\mu_\perp^\star = 0$. Hence $\mu^\star(x) = \rho\nu(x)u$.

Optimal variance. With $\sigma_r^2(x)$ the r-th coordinate of $\Sigma_\theta(x)$, each coordinate contributes

$$\frac{u_r^2}{\rho^2}\sigma_r^2 + \frac{\beta}{2}\Big[\frac{\sigma_r^2}{\tau^2} - 1 - \log\frac{\sigma_r^2}{\tau^2}\Big],$$

whose derivative in σ_r^2 is $\frac{u_r^2}{\rho^2} + \frac{\beta}{2}\big(\frac{1}{\tau^2} - \frac{1}{\sigma_r^2}\big)$ and whose second derivative is $\frac{\beta}{2\sigma_r^4} > 0$; the unique minimizer is

$$\sigma_r^{2\star} = \Big(\frac{1}{\tau^2} + \frac{2u_r^2}{\beta\rho^2}\Big)^{-1} \le \tau^2, \tag{26}$$

with equality exactly when $u_r = 0$, tending to τ^2 as $\beta \to \infty$ or $\rho \to \infty$.

Mean matching only. Without the variance KL, the variance enters the objective only through $\frac{1}{\rho^2}\mathbb{E}[u^\top\Sigma_\theta u]$, whose pointwise infimum over σ_r^2 is approached only as $\sigma_r^2 \to 0$ whenever $u_r \neq 0$: pure mean matching drives readout-direction noise to zero, while the full KL selects the interior balance of Eq. (26). The same mechanism holds for classification: $\mathrm{CE}(\eta, \mathrm{softmax}(\cdot))$ is convex in the logits, so Jensen's inequality gives $\mathbb{E}_\xi\,\mathrm{CE}(\eta(x), p_\mu(\xi; x)) \ge \mathrm{CE}(\eta(x), p_\mu(0; x))$ and the sampled-path risk likewise favors noise-free readout directions, resisted only by the variance KL (the GPB-M ablation of Appendix A.4.4). Both objectives attain the Bayes mean in this ideal setting, so the balance mechanism explains the variance regularization's role without by itself predicting test-performance differences.

Predictive consequence of approximate anchor matching. For a feasible model w with excess mismatch $A_m(w)$ (Eq. (14)), let $\mathcal{E}_{\mathrm{cls}}(g) = P(g(X) \neq Y) - \inf_{g'} P(g'(X) \neq Y)$ and $\mathcal{E}_{\mathrm{reg}}(f) = \mathbb{E}[(f(X) - \tilde{Y})^2] - \inf_{f'}\mathbb{E}[(f'(X) - \tilde{Y})^2]$ be the excess risks against the Bayes predictors. The OPB mean readout $g_\mu(x) = \arg\max_k q_{\theta,k}^\top \mu_\phi(x)$ (the anchor-energy classifier at $z = \mu_\phi(x)$ with uniform score priors) and the EPB mean readout $f_\mu(x) = u^\top\mu_\phi(x)/\rho$ then satisfy

$$\mathcal{E}_{\mathrm{cls}}(g_\mu) \le \frac{2\tau}{a}\sqrt{A_m(w)}, \qquad \mathcal{E}_{\mathrm{reg}}(f_\mu) \le \frac{2\tau^2}{\rho^2}A_m(w), \tag{27}$$

so approximate anchor matching suffices for approximately Bayes-optimal prediction. The classification bound follows from the plug-in regret argument (a score vector v with $\arg\max v \neq \arg\max \eta$ loses at most $\sqrt{2}\|v - \eta\|_2$ label probability) and the contraction of $Q_\theta^\top$; the regression bound from the squared-loss excess-risk identity and the projection onto u.

Table 5: Per-method hyperparameter grids of the main comparison (Table 1); the best validation configuration is selected per method and setting.

Method	Method-specific hyperparameters	Configs.
Base	—	1
VIB, SVIB, NIB, CEB, DVCCA	$\beta \in \{1, 2, 4, 6, 8\} \times 10^k$, $k = -4, \ldots, 1$	30
NCBD	temperature T: 30 log-spaced values in $[0.05, 1]$	30
AdaCap	Tikhonov $\lambda \in \{1, 2, 4, 6, 8\} \times 10^k$, $k = -2, \ldots, 3$	30
GPB (OPB / EPB)	$\beta \in \{10^{-4}, \ldots, 10\}$; $a = \rho \in \{1, 3, 6, 9, 12\}$	$6 \times 5 = 30$

A.3 Setup and construction details

A.3.1 Complete experimental setup

Scope and shared settings. We evaluate twelve dataset–backbone settings over nine public datasets: MNIST, ImageNet-100, Cora, IMDb, AG News, California Housing, STS-B, ZINC-12k, and AgeDB (Lecun et al., 1998; Deng et al., 2009; Sen et al., 2008; Maas et al., 2011; Zhang et al., 2015; Kelley Pace & Barry, 1997; Cer et al., 2017; Irwin et al., 2012; Moschoglou et al., 2017), covering MLP, CNN, GCN, and RNN backbones. All methods use the same data splits ($60/20/20$ train/validation/test with a fixed seed for most datasets, official splits kept where provided, e.g., Cora and ZINC-12k), backbone, Gaussian posterior with learned diagonal variance, bottleneck $d = 256$, Adam (Kingma & Ba, 2017) training, early stopping, and validation-based selection; GPB's prior covariance is the fixed shared $\tau^2 I$ ($\tau^2 = 1$). Stochastic bottleneck methods are trained with reparameterized samples from their encoder posteriors, while validation, checkpoint selection, and test evaluation use posterior-mean inference: the encoder mean is passed to each method's own prediction head (the anchor-energy classifier for OPB, the tied projection for EPB). Deterministic methods use their ordinary prediction rule. Sampled-path predictions are used only where stated: the robustness study (MC-10/EOT-10, Sec. 3.4) and the information-plane estimates ($z \sim q(z|x)$, Sec. 3.8).

Training defaults and preprocessing. The shared training defaults are hidden sizes 512–256, dropout 0.2, learning rate 10^{-3} (Cora uses 10^{-2}), batch size 512, and a 200-epoch budget with early stopping (patience 20); dataset-specific defaults are likewise shared by every method on that dataset: STS-B uses a single-layer LSTM with hidden size 256 and frozen GloVe (Pennington et al., 2014) embeddings; AG News uses 50-dimensional PCA-reduced BERT (Devlin et al., 2019) features; ImageNet-100 uses 4×4 spatial pool-4 ResNet-50 (He et al., 2015) features; the remaining dataset-specific preprocessing is given in the release. Every configuration is trained on the same ten seeds (0–9). Input features are standardized per dataset, and regression targets are min–max normalized to $[0, 1]$; both are fitted on the training split only, and reported MAE is rescaled to the original units.

Baselines and search-budget fairness. The baselines are Base, VIB, SVIB, NIB, CEB, supervised β-DVCCA, and the external non-IB baselines NCBD and AdaCap: NCBD trains unit-hypersphere features and prototypes with the NONL contrastive objective (simplex-ETF optimum), and AdaCap pairs a permutation contrastive loss with a Tikhonov closed-form readout, so the comparison spans geometric/prototype and contrastive non-IB mechanisms in addition to the IB variants. GPB is OPB on classification rows and EPB on regression rows. The main comparison equalizes the search budget: the plain backbone has no method-specific hyperparameters (one configuration); every β-weighted method, VIB, SVIB, NIB, CEB, and DVCCA, is swept over the same 30-point β grid (six decades, five values per decade); GPB searches the two-dimensional grid $\beta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\} \times a = \rho \in \{1, 3, 6, 9, 12\}$, also $6 \times 5 = 30$ configurations; NCBD and AdaCap use 30-value grids of their own. The same validation-based selection rule applies to every ablation variant (Appendix A.4.4). The per-method grids are summarized in Table 5.

A.3.2 Construction details

OPB construction notes. The construction is equivariant under label permutations ($Q_\theta(U_\theta P) = Q_\theta(U_\theta)P$), so no class index plays a distinguished role, and the prior encoder is identity-initialized, so the initial raw matrix $U_\theta = [e_1, \ldots, e_K]$ is full column rank with a stable polar-factor backward. The prior covariance is the fixed $\tau^2 I$ with $\tau^2 = 1$ and no learned parameters. For uniform classes

Table 6: Test accuracy (%) and test R^2 with standard deviations over runs: Base, VIB, SVIB, NIB, and NCBD (classification settings only); same configurations as Table 1.

Dataset	Backbone	Base	VIB	SVIB	NIB	NCBD
MNIST	MLP	98.33±0.12	98.28±0.11	98.15±0.13	98.27±0.07	97.95±0.23
MNIST	CNN	99.13±0.05	99.28±0.09	99.23±0.03	99.22±0.06	99.18±0.07
ImageNet-100	MLP	83.12±0.14	82.84±0.44	83.02±0.37	82.35±0.27	83.64±0.07
ImageNet-100	CNN	83.40±0.13	82.94±0.37	83.11±0.86	82.91±0.35	80.56±1.86
Cora	GCN	87.68±0.46	86.83±0.28	86.46±1.87	86.35±0.68	87.34±1.60
IMDb	RNN	84.85±0.68	85.50±0.91	85.99±0.30	86.14±0.92	86.69±0.52
AG News	RNN	**92.94±0.15**	92.72±0.17	92.86±0.21	92.91±0.22	92.37±0.37
Cal. Housing	MLP	0.7618±0.0238	**0.7901±0.0049**	0.7897±0.0058	0.7862±0.0035	–
STS-B	RNN	**0.2899±0.0113**	0.2719±0.0164	0.2697±0.0202	0.2417±0.0221	–
ZINC	GCN	0.6348±0.0044	0.6339±0.0033	0.6325±0.0026	0.6178±0.0099	–
AgeDB	MLP	0.2180±0.0409	0.3173±0.0101	0.3385±0.0286	0.2710±0.0235	–
AgeDB	CNN	0.3914±0.0194	0.3342±0.1894	0.3323±0.1870	0.3337±0.1889	–

Table 7: Test accuracy (%) and test R^2 with standard deviations over runs: CEB, DVCCA, AdaCap (regression settings only), and GPB; same configurations as Table 1.

Dataset	Backbone	CEB	DVCCA	AdaCap	GPB
MNIST	MLP	98.43±0.08	98.24±0.17	–	**98.54±0.09**
MNIST	CNN	99.30±0.05	99.29±0.10	–	**99.32±0.05**
ImageNet-100	MLP	82.77±0.22	82.99±0.43	–	**84.40±0.21**
ImageNet-100	CNN	82.97±0.67	83.21±0.70	–	**86.08±0.21**
Cora	GCN	86.27±0.38	86.90±0.94	–	**88.12±0.55**
IMDb	RNN	85.55±0.73	85.69±0.23	–	**86.84±0.57**
AG News	RNN	92.83±0.11	92.78±0.15	–	92.87±0.12
Cal. Housing	MLP	0.7894±0.0018	0.7878±0.0048	0.7758±0.0023	0.7895±0.0051
STS-B	RNN	0.2668±0.0127	0.2713±0.0218	0.1613±0.0188	0.2762±0.0152
ZINC	GCN	0.6257±0.0044	0.6294±0.0026	**0.6770±0.0019**	0.6344±0.0031
AgeDB	MLP	0.3317±0.0251	0.3180±0.0228	0.3482±0.0209	**0.3536±0.0076**
AgeDB	CNN	0.3882±0.0275	0.3205±0.1811	0.4684±0.0094	**0.4700±0.0072**

the anchor-energy scores of Eq. (10) are linear in z up to a class-independent term, with weights $(a/\tau^2)q_{\theta,k}$, so the readout scale is fixed by the prior alone.

EPB isometry notes. The exactness is automatic: any nonzero linear map of a scalar label is isometric up to its norm, and the normalization fixes that norm to ρ. All prior means lie on one line through the origin, so anchor norms grow with $|\tilde{y}|$: a line and a fixed-radius sphere intersect in at most two points, so isometry and equal norms cannot both hold. For $y \in \mathbb{R}^m$ ($m \leq d$) the same construction extends verbatim: take $W \in \mathbb{R}^{d\times m}$ and its thin polar factor Q_θ as in Eq. (8) with K replaced by m, and the prior $m_\theta(y) = \rho\, Q_\theta\, \tilde{y}$, isometric in the declared metric on the normalized labels, the scalar case being $m = 1$ with $Q_\theta = u$. The label metric must be declared before normalization, otherwise "isometric" has no unique meaning.

A.4 Additional experiments

A.4.1 Main results: standard deviations and confidence intervals

Tables 6 and 7 report the same configurations' metrics as the main table, with standard deviations over runs, split by column groups for width; bold and underline follow the same mean-based rule as Table 1.

Tables 8–10 report, for each setting, the paired difference of GPB over every baseline (GPB minus baseline) with a bootstrap 95% confidence interval: the differences are paired across the shared seeds, the interval is the percentile interval (Efron & Tibshirani, 1994) of $B = 10{,}000$ resamples of the per-run differences (fixed seed, reproducible), and the non-inferiority margin is $\delta_{\mathrm{NI}} = 0.2$ percentage points of accuracy or $\delta_{\mathrm{NI}} = 0.005$ in R^2. Bold marks a lower bound above zero (significant advantage) and † a lower bound within the margin (tied); the best configuration per method is the

Table 8: Paired-difference 95% confidence intervals of GPB over Base and VIB; GPB minus baseline, per setting; classification in percentage points, regression in R^2. Bold: lower CI bound above zero; †: lower bound within the non-inferiority margin δ_{NI} (tied).

Dataset	Backbone	Base	VIB
MNIST	MLP	**+0.2 [+0.2, +0.3]**	**+0.3 [+0.2, +0.3]**
MNIST	CNN	**+0.2 [+0.2, +0.2]**	+0.0 [-0.0, +0.1]†
ImageNet-100	MLP	**+1.3 [+1.2, +1.4]**	**+1.6 [+1.4, +1.9]**
ImageNet-100	CNN	**+2.7 [+2.5, +2.9]**	**+3.1 [+3.0, +3.4]**
Cora	GCN	+0.4 [-0.1, +1.1]†	**+1.3 [+0.7, +1.7]**
IMDb	RNN	**+2.0 [+1.2, +2.7]**	**+1.3 [+0.6, +2.3]**
AG News	RNN	-0.1 [-0.2, +0.0]†	**+0.2 [+0.1, +0.2]**
Cal. Housing	MLP	**+0.028 [+0.008, +0.048]**	-0.001 [-0.005, +0.003]†
STS-B	RNN	-0.014 [-0.022, -0.006]	+0.004 [-0.017, +0.022]
ZINC	GCN	-0.000 [-0.003, +0.002]†	+0.001 [-0.004, +0.004]†
AgeDB	MLP	**+0.136 [+0.104, +0.167]**	**+0.036 [+0.026, +0.046]**
AgeDB	CNN	**+0.079 [+0.066, +0.091]**	**+0.136 [+0.043, +0.302]**
		8*/3†	8*/3†

Table 9: Paired-difference 95% confidence intervals of GPB over SVIB, NIB, and CEB; GPB minus baseline, per setting; classification in percentage points, regression in R^2. Bold: lower CI bound above zero; †: lower bound within the non-inferiority margin δ_{NI} (tied).

Dataset	Backbone	SVIB	NIB	CEB
MNIST	MLP	**+0.4 [+0.3, +0.5]**	**+0.3 [+0.2, +0.3]**	**+0.1 [+0.0, +0.2]**
MNIST	CNN	**+0.1 [+0.1, +0.1]**	**+0.1 [+0.1, +0.2]**	+0.0 [-0.0, +0.1]†
ImageNet-100	MLP	**+1.4 [+1.1, +1.7]**	**+2.1 [+1.8, +2.2]**	**+1.6 [+1.5, +1.7]**
ImageNet-100	CNN	**+3.0 [+2.4, +3.9]**	**+3.2 [+2.9, +3.5]**	**+3.1 [+2.4, +3.7]**
Cora	GCN	**+1.7 [+0.2, +3.4]**	**+1.8 [+1.1, +2.4]**	**+1.8 [+1.5, +2.3]**
IMDb	RNN	**+0.8 [+0.2, +1.4]**	**+0.7 [+0.2, +1.2]**	**+1.3 [+0.3, +2.1]**
AG News	RNN	+0.0 [-0.1, +0.1]†	-0.0 [-0.2, +0.1]†	+0.0 [-0.1, +0.2]†
Cal. Housing	MLP	-0.000 [-0.006, +0.004]	+0.003 [-0.001, +0.009]†	+0.000 [-0.003, +0.004]†
STS-B	RNN	+0.006 [-0.021, +0.028]	**+0.035 [+0.011, +0.058]**	+0.009 [-0.009, +0.025]
ZINC	GCN	+0.002 [-0.001, +0.004]†	**+0.017 [+0.010, +0.026]**	**+0.009 [+0.005, +0.011]**
AgeDB	MLP	+0.015 [-0.010, +0.041]	**+0.083 [+0.062, +0.100]**	**+0.022 [+0.006, +0.038]**
AgeDB	CNN	**+0.138 [+0.045, +0.303]**	**+0.136 [+0.040, +0.302]**	**+0.082 [+0.064, +0.099]**
		7*/2†	10*/2†	8*/3†

same as in the main table. The last column of Table 1 counts, per setting, the number of baselines GPB significantly outperforms versus is significantly outperformed by, under the same paired confidence intervals. For the aggregate counts in Section 3.2, a Holm–Bonferroni correction at $\alpha = 0.05$ is applied over the 84 applicable paired comparisons, with two-sided p-values derived from the same bootstrap distributions.

A.4.2 Controlled objective selection: protocol and details

The controlled study of Section 3.3 numerically minimizes two population objectives over a shared encoder family on a distribution whose label posterior is known analytically. The input is $X \sim \mathrm{Bernoulli}(1/2)$ and the label channel is binary symmetric, $Y = X \oplus N$ with $N \sim \mathrm{Bernoulli}(r)$, $r \in \{0.1, 0.2, 0.4\}$. The latent space is two-dimensional with $Q = I_2$, $\tau^2 = 1$, posterior covariance $\Sigma = I_2$ fixed, and OPB anchors $m_0 = ae_0$, $m_1 = ae_1$ with $a = 2$. The shared encoder family is

$$q_t(z|x) = \mathcal{N}\big(\tilde{c} + (2x-1)t\tilde{v},\ I_2\big), \qquad \tilde{c} = \frac{a}{2}(e_0 + e_1), \quad \tilde{v} = \frac{e_1 - e_0}{\sqrt{2}}, \tag{28}$$

parameterized by a scalar $t \geq 0$; $t = 0$ is input-independent, and in the projected coordinate $\tilde{s} = \tilde{v}^\top(z - \tilde{c})$ the two conditionals are $q_t(\tilde{s}|x = 0) = \mathcal{N}(-t, 1)$ and $q_t(\tilde{s}|x = 1) = \mathcal{N}(t, 1)$.

CEB-oracle. The idealized CEB objective uses the exact aggregated reference and Bayes readout, so

$$\mathcal{L}^\star_{\mathrm{CEB}}(t) = H_t(Y|Z) + \beta I_t(X; Z|Y) = \ln 2 + \beta I_t(X; Z) - (1 + \beta) I_t(Y; Z), \tag{29}$$

Table 10: Paired-difference 95% confidence intervals of GPB over DVCCA, NCBD (classification settings only), and AdaCap (regression settings only); GPB minus baseline, per setting; classification in percentage points, regression in R^2. Bold: lower CI bound above zero; †: lower bound within the non-inferiority margin δ_{NI} (tied).

Dataset	Backbone	DVCCA	NCBD	AdaCap
MNIST	MLP	**+0.3 [+0.2, +0.4]**	**+0.6 [+0.4, +0.8]**	–
MNIST	CNN	+0.0 [-0.1, +0.1]†	**+0.1 [+0.1, +0.2]**	–
ImageNet-100	MLP	**+1.4 [+1.1, +1.7]**	**+0.8 [+0.6, +0.9]**	–
ImageNet-100	CNN	**+2.9 [+2.4, +3.5]**	**+5.5 [+4.3, +6.9]**	–
Cora	GCN	**+1.2 [+0.4, +1.9]**	**+0.8 [+0.0, +1.9]**	–
IMDb	RNN	**+1.1 [+0.8, +1.5]**	+0.2 [-0.5, +0.7]	–
AG News	RNN	+0.1 [-0.0, +0.2]†	**+0.5 [+0.2, +0.8]**	–
Cal. Housing	MLP	+0.002 [-0.005, +0.009]	–	**+0.014 [+0.009, +0.019]**
STS-B	RNN	+0.005 [-0.020, +0.026]	–	**+0.115 [+0.086, +0.135]**
ZINC	GCN	**+0.005 [+0.002, +0.007]**	–	-0.043 [-0.046, -0.041]
AgeDB	MLP	**+0.036 [+0.012, +0.057]**	–	+0.005 [-0.007, +0.021]
AgeDB	CNN	**+0.149 [+0.065, +0.310]**	–	+0.002 [-0.004, +0.006]†
		8*/2†	6*/0†	2*/1†

where $q_t(s)$ is the uniform two-Gaussian mixture and $q_t(s|y)$ the mixture weighted by the noise rate, and the information terms are computed by one-dimensional Gauss–Hermite quadrature; no mutual-information critic and no Gaussian approximation of the mixture is used.

OPB. The OPB objective is implemented directly with the declared prior $\mathcal{N}(m_y, I_2)$ and the tied energy readout under uniform class priors. Two closed forms keep it one-dimensional: the energy logit difference is $a\sqrt{2}\,\tilde{s}$ (the component orthogonal to $\tilde{v}$ cancels), and the KL mean term is

$$\mathbb{E}\,\mathrm{KL}\big(q_t(Z|X)\,\|\,p(Z|Y)\big) = \frac{1}{2}\Big[(1-r)\Big(\frac{a}{\sqrt{2}}-t\Big)^2 + r\Big(\frac{a}{\sqrt{2}}+t\Big)^2\Big], \tag{30}$$

its variance term vanishing since $\Sigma = \tau^2 I$.

Numerical protocol. The β grid is $\mathrm{logspace}(-3, 3, 49)$ plus points bracketing the threshold $\beta_{\mathrm{SDPI}} = (1-2r)^2/(4r(1-r))$; for each β the scalar t is solved by coarse grid search (including the explicit candidate $t = 0$), local bounded refinement, and adaptive expansion of the upper bound until inactive. Gauss–Hermite quadrature uses $n = 128$ nodes; doubling to 256 changes every reported objective by at most 1.9×10^{-7}, and the information-theoretic integrals agree with adaptive quadrature to 10^{-9}. The reference values are the conditional-expectation target $t_\infty = a(1-2r)/\sqrt{2}$, the mean-path error $R_{\mathrm{mean}} = r$ for $t > 0$, and the sampled-path error $R_{\mathrm{sample}} = r + (1-2r)\Phi(-t)$. This is a population-objective experiment on a controlled encoder family, not a numerical search over neural-network optima.

Off-figure observations. The measured CEB transition points match the closed-form threshold $\beta_{\mathrm{SDPI}} = (1-2r)^2/(4r(1-r))$ at every noise rate (1.778, 0.5625, 0.04167 for $r = 0.1, 0.2, 0.4$); the OPB amplitude at $\beta = 10^3$ reads 0.8486 against $t_\infty = 0.8485$ at $r = 0.2$ (similarly 1.1316 at $r = 0.1$ and 0.2829 at $r = 0.4$), and $I(X;Z)$ plateaus at 0.267 at $r = 0.2$. At weak compression, $t^\star$ lies slightly above t_∞, near the finite minimizer of the sampled cross-entropy induced by the fixed-scale energy readout. On the sampled path the error $r + (1-2r)\Phi(-t)$ rises from the Bayes error r toward $1/2$ as t shrinks (0.319 at $\beta = 10^3$, $r = 0.2$): mean-path Bayes retention does not certify a non-collapsed sampled representation.

A.4.3 Adversarial robustness details

This appendix supplements Section 3.4, which measures targeted white-box PGD success on MNIST, with the protocol details and a partial gradient-masking check. Targeted PGD uses $T = 20$ steps, EOT-10 gradients, 10-sample MC predictions, random starts, step size 2.5ε divided by the number of PGD steps, and budgets $\varepsilon_\infty = 0.3$ and $\varepsilon_2 = 6$; the target is always class 0, and success is counted only on clean-correct test examples whose label differs from the target class. For the transfer check, adversaries generated on one model are evaluated on another (CEB to OPB and back); their transferred success (Fig. 5) stays far below the white-box level of Fig. 2 at every budget. This

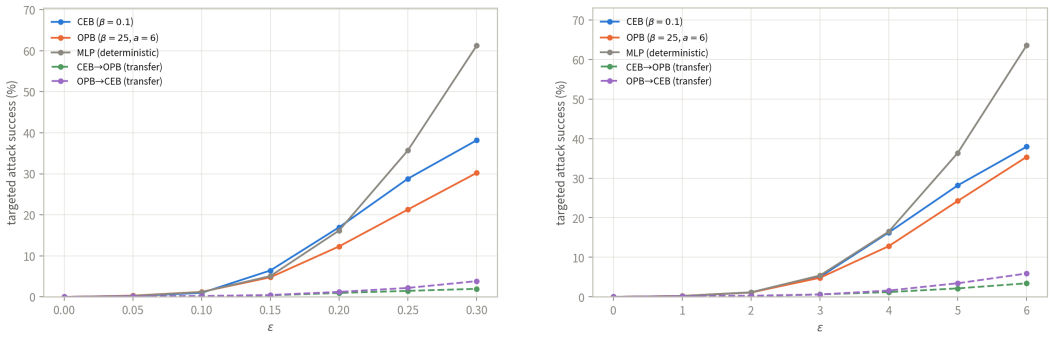

Figure 5: Transfer PGD curves on MNIST (partial gradient-masking check). Targeted adversaries are generated on one model and evaluated on another. The targeted white-box curves are in the main text.

Table 11: Structural anatomy of the ablation variants: each variant switches one aspect of the full model.

variant	posterior	penalty	frame	initialization	variance	readout
GPB	stochastic	KL	trainable polar	identity	fixed τ^2=1	energy
GPB-H	stochastic	KL	trainable polar	identity	fixed τ^2=1	free head
GPB-V	stochastic	KL	trainable polar	identity	learnable	energy
GPB-P	stochastic	KL	fixed polar	identity	fixed τ^2=1	energy
NCM-G	deterministic	none	fixed polar	fixed	fixed τ^2=1	energy
GPB-G	stochastic	KL	free anchors	identity	fixed τ^2=1	energy
NCM-T	deterministic	none	trainable polar	identity	fixed τ^2=1	energy
NCM-S	stochastic	none	trainable polar	identity	fixed τ^2=1	energy
NCM-M	deterministic	mean matching	trainable polar	identity	fixed τ^2=1	energy
GPB-M	stochastic	mean matching	trainable polar	identity	fixed τ^2=1	energy
GPB-CS	stochastic	KL	trainable ETF	identity	fixed τ^2=1	energy

shows that the adversarial directions are model-specific: perturbations found through one model's gradients barely affect the other, so the check finds no evidence of the masking signature in which perturbations transfer successfully while white-box optimization fails; it does not exclude gradient masking.

A.4.4 Item-wise ablations

The item-wise ablations switch the prior geometry, the geometry convention, the readout, the prior variance, frame trainability, and the IB term separately; every variant is evaluated at its best validation-selected configuration. The six factor variants are described in the main text (Section 3.5), the four decomposition variants below are defined where they are introduced, and the structural anatomy of all variants is summarized in Table 11, whose columns read as follows. *Posterior* indicates whether z is sampled from $q(z|x)$ (stochastic) or fixed at the encoder mean μ (deterministic). *Penalty* is the regularization term beyond the cross-entropy: the conditional KL, the mean-matching term $\frac{\beta}{2\tau^2}\|\mu(x) - m_y\|_2^2$ weighted by β, i.e., the exact mean part of the conditional KL with the variance term removed, or none. *Frame* is the geometry of the prior means: the trainable polar frame of GPB (trainable polar), the same frame frozen at the identity anchors $a \cdot [e_1..e_K]$ (fixed polar), class-specific anchor means without the orthogonality constraint (free anchors), or the trainable polar frame centered at every forward pass into a simplex table (trainable ETF). *Initialization* is the starting anchor configuration: the identity frame $a \cdot [e_1..e_K]$ (identity), or fixed constants that are never updated (fixed). *Variance* is the prior covariance τ^2, fixed at 1 (τ^2=1) or learned per class (learnable). *Readout* is the prediction head: the anchor-energy classifier of GPB (energy) or a free linear head (free head). The paired differences with bootstrap intervals and the variant's mean for the five factor contrasts are in the main text (Table 2); the finer IB-increment decomposition below (Table 12) follows the same paired percentile protocol as Section A.4.1. The factor ranking is read in the main text (Section 3.5).

Table 12: Decomposition of the IB increment on ImageNet-100 (MLP): for each contrast, the paired difference in accuracy points with its 95% bootstrap percentile CI and the variant's mean ± standard deviation over seeds.

comparison	difference [95% CI]	variant mean ± std
GPB − NCM-G	+2.07 [1.59, 2.55]	82.33 ± 0.60
GPB − NCM-S	+2.43 [2.09, 2.82]	81.98 ± 0.52
GPB − NCM-M	+2.56 [1.96, 3.20]	81.84 ± 0.75
GPB − GPB-M	+1.84 [1.58, 2.11]	82.56 ± 0.16

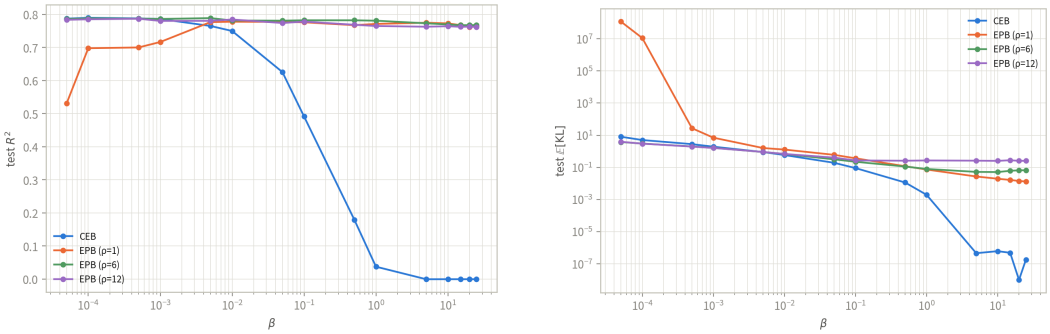

Figure 6: California Housing compression sweep (MLP). Left: test R^2; right: realized conditional KL over β.

How large is the IB increment once the trainable frame is matched on both sides? Removing the IB term while keeping the trainable polar frame and the energy readout (NCM-T: deterministic output, no KL) costs 2.69 points, the increment reported in the main table. The fixed-frame contrast (GPB − NCM-G: +2.07) understates it: frame trainability is at best inert and at worst mildly harmful in the deterministic setting (NCM-T − NCM-G: −0.62), so its value is tied to the conditional KL and the trainable-frame control is the fair reference for GPB.

Does the benefit come from the injected noise itself rather than from the KL? Restoring the stochastic encoder without the KL (NCM-S, noise injection only) recovers merely 0.26 of the 2.69 points (GPB − NCM-S: 2.43). The KL term therefore carries most of the IB increment, while noise injection alone is nearly inert.

Is the mechanism reducible to supervised anchor regression, i.e., pulling posterior means toward class anchors? Replacing the KL by pure mean matching in the deterministic model (NCM-M, loss CE + $\frac{\beta}{2\tau^2}\|\mu(x) - m_y\|_2^2$) leaves a gap of 2.56 points. Mean alignment recovers almost none of the increment, ruling out supervised anchor regression as its source.

Does the variance and noise regularization inside the KL contribute beyond mean matching? Keeping the stochastic encoder but replacing the KL by pure mean matching (GPB-M, loss CE + $\frac{\beta}{2\tau^2}\|\mu(x) - m_y\|_2^2$) loses 1.84 points, so the variance regularization has a significant contribution of its own beyond mean matching, and the gain is not merely supervised anchor regression.

A.4.5 Regression compression curves

Section 3.7 shows the compression sweep on classification; this section shows the regression counterpart on California Housing under the same β grid. Fig. 6 shows: EPB maintains its test-R^2 plateau through strong compression, and only the $\rho = 1$ weak-compression points exhibit a large variance term, so the regression side reproduces the classification behavior.

A.4.6 Anchor-scale sensitivity

Fig. 7 reads the full nine-value scale grid $\{1, 2, 4, 6, 8, 10, 12, 14, 16\}$ at weak, medium, and strong compression on the two representative settings, complementing the β-axis sweeps of Sec. 3.7 with the scale-axis view.

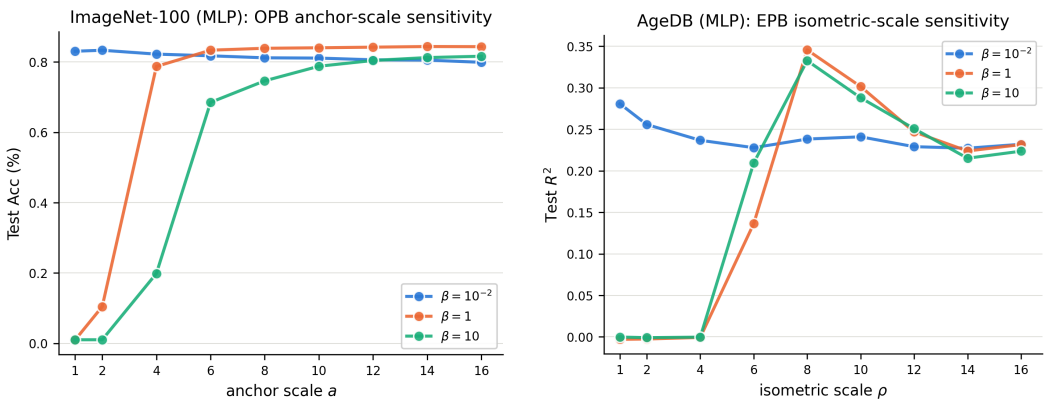

Figure 7: Anchor-scale sensitivity: (a) ImageNet-100 (MLP) OPB test accuracy against the anchor scale a; (b) AgeDB (MLP) EPB test R^2 against the isometric scale ρ; $\beta \in \{10^{-2}, 1, 10\}$ (weak/medium/strong compression).

On both settings the scale matters only once compression bites: the weak-compression curves are flat, and at $\beta \geq 1$ the smallest scales collapse (chance accuracy on ImageNet-100, zero R^2 on AgeDB), while scales above 6 stay well above the collapse boundary. The curves are not flat beyond that boundary, however: on AgeDB, R^2 peaks near $\rho = 6$–8 (≈ 0.36) and declines to about 0.21 at the largest scales, and ImageNet-100 at strong compression ramps up with a before a late dip. The operating point is therefore sensitive to the scale mainly through the collapse boundary, with a secondary non-monotone dependence beyond it.

A.4.7 Information-plane estimates

Estimator protocol. $I(X;Z)$ is estimated by the InfoNCE lower bound (van den Oord et al., 2019; Poole et al., 2019) ($\log N_{\text{neg}} - \mathcal{L}_{\text{NCE}}$, $N_{\text{neg}} = 4096$; one critic per run, trained on the training split, the bound read off the test split with $z \sim q(z|x)$). The critic is the CPC-style bilinear score $s = g_z(z)^\top g_h(h)$ with two 128-unit MLP heads (Linear–ReLU–Linear); its x-side input is the encoder output h, which loses no information on the μ path since $z = \text{Linear}(h)$. The critic is trained with Adam (learning rate 10^{-3}) on batches of 4096 for an adaptive budget of roughly 30 epochs, and its training loss plateaus well within that budget. $I(Y;Z)$ is read on the same sampled posterior: $H(Y) - \text{CE}$ with the cross-entropy averaged over $M = 20$ draws $z \sim q(z|x)$ per test point (still a lower bound, as $\text{CE} \geq H(Y|Z)$). The residual $I(X;Z|Y)$ is read on the known Gaussian posteriors: the class-conditional mixture $q(z|y) = \frac{1}{N_y} \sum_{j:y_j=y} q(z|x_j)$ over the test samples is evaluated exactly by logsumexp, and $I(X;Z|Y) = \mathbb{E}_{x,y} \, \text{KL}(q(z|x) \,\|\, q(z|y))$ is read by Monte Carlo with $M = 20$ z-draws per test point. The mixture includes the query sample itself (self-inclusion weight $1/N_y \approx 10^{-3}$ per class on MNIST, a negligible bias here). The certificate-plane axis is drawn linear because the collapsed CEB trajectory lands on zero.

Matched lower-bound comparison. Table 13 pairs each CEB point with the OPB configuration of nearest InfoNCE lower bound $\widehat{I}_{\text{NCE}}(X;Z)$: at the first matched level OPB is within 0.1 point of CEB with a comparable residual, and at the second level OPB keeps 97.8 against CEB's 86.0, so OPB's task plateau is not bought by extra residual compression at equal estimated compression. Matching estimated lower bounds does not establish equal true compression, since the critic's approximation quality can differ across models and β (Appendix A.4.7); these are paired empirical comparisons at two matched lower-bound levels, not a causal attribution to geometry.

Regression companion. Figure 8 verifies on the regression setting what the classification panels of Section 3.8 establish, namely what each method compresses and discards as compression deepens; here $I(Y;Z)$ is the Gaussian approximation $\frac{1}{2} \ln\big(\text{Var}(y)/\text{Var}(y - \hat{y}_{\text{MC}})\big)$ read on the sampled posterior, and the residual $I(X;Z|Y)$ uses a Gaussian-kernel-weighted mixture of the test posteriors ($\hat{q}(z|y) = \sum_j w_j(y) \, q(z|x_j)$, with the bandwidth set by Silverman's rule on the training split), evaluated exactly by logsumexp; every term is a KL between proper densities, so the estimate is

Table 13: Matched InfoNCE lower-bound comparison. The OPB configuration is the one whose estimated $\widehat{I}_{\mathrm{NCE}}(X;Z)$ is nearest to the CEB point.

	CEB			OPB		
Setting	$I(X;Z)$	$I(X;Z\|Y)$	Acc.	$I(X;Z)$	$I(X;Z\|Y)$	Acc.
MNIST	2.54	0.41	98.2	2.55	0.47	98.1
MNIST	1.96	0.20	86.0	2.01	0.21	97.8

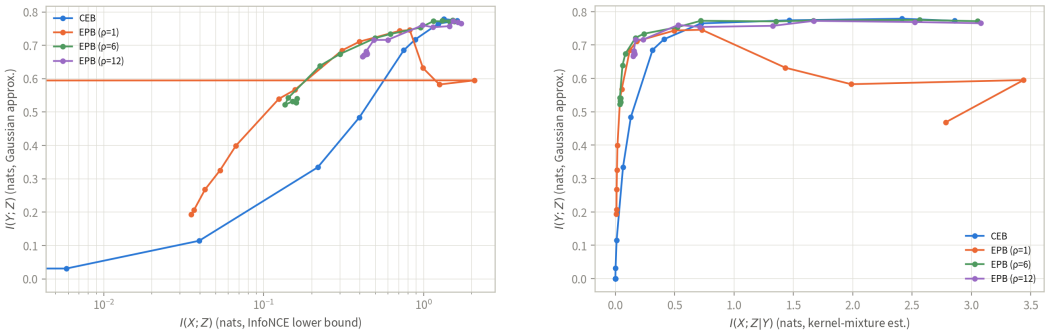

Figure 8: Information plane (a) and certificate plane (b) for Cal. Housing, companion panels to Fig. 4(c,d).

non-negative by construction. The two panels of Fig. 8 show the same phenomenon: CEB's residual and task information fall toward zero together, whereas EPB ends strong compression on a small positive residual plateau whose task-information level grows with the scale; every EPB curve keeps its R^2 flat out to the strongest compression, so compression again stops where the task information sits.

A.5 Related work

Information bottleneck and conditional compression. The information bottleneck (IB) principle formulates representation learning as retaining information about a target while discarding information about the input (Tishby et al., 2000). Deep VIB turns this principle into a neural objective by using a reparameterized stochastic encoder and a variational marginal prior (Alemi et al., 2019). Nonlinear and non-parametric bottleneck variants study alternative variational bounds and distance-based estimates of the information in a representation (Kolchinsky et al., 2019b). The deterministic-scenario analysis of Kolchinsky et al. (2019a) further shows that the usual IB multiplier can fail to sweep an interior information–accuracy trade-off when labels are deterministic, motivating the squared-IB functional, which squares the compression term and restores the interior trade-off.

The conditional entropy bottleneck (CEB) replaces the marginal reference of VIB by a label-conditional reference and thereby targets conditional residual information (Fischer, 2020). Comparisons of variational bounds clarify when this conditional reference is tighter than a marginal one (Geiger & Fischer, 2020). CEB has also been applied to large-scale visual representations and robustness evaluations (Fischer & Alemi, 2020; Lee et al., 2021), while multivariate extensions retain the same variational bottleneck viewpoint (Abdelaleem et al., 2025), and regression-side IB bounds provide compression objectives without a geometric prior construction (Yu et al., 2024). These methods leave the relationships among distinct conditional reference means unspecified; Sec. 2.1 formalizes this freedom as a degeneracy specific to unconstrained label-conditioned references. GPB keeps the CEB posterior and its single conditional KL, but restricts the reference family itself: OPB declares a class frame and EPB declares a label-distance-preserving axis.

Information bottleneck for robust representations. Several works use a bottleneck to suppress nuisance or non-robust information. The HSIC bottleneck adds layer-wise dependence penalties and analyzes adversarial robustness (Wang et al., 2021). Adversarial Information Bottleneck explicitly incorporates adversarial objectives (Zhai & Zhang, 2024), and an NLP-specific variant separates

task-relevant robust features from manipulable features (Zhang et al., 2022). Causal Information Bottleneck uses a causal decomposition for the same robust-feature goal (Hua et al., 2022). These approaches modify the information criterion, feature selection, or training examples. In contrast, the geometric constraint in GPB is placed on the conditional prior and is evaluated with the ordinary CEB KL, without any additional objective; under the evaluated targeted-PGD protocol of Appendix A.4.3, the compressed geometric operating point shows lower attack success than CEB's last valid operating point, subject to the masking caveats discussed there.

Geometric representations. Neural-collapse studies document an approximately simplex-ETF arrangement of class means and classifier directions (Papyan et al., 2020; Zhu et al., 2021). Other works impose related geometry directly: orthogonal classifier weights (Xu et al., 2022), ETF classifiers for early exit (Ji et al., 2023), and fixed or hierarchy-aware frames (Liang & Davis, 2023). Prototype-based constructions date back to equidistant prototype embeddings (Deng et al., 2014); Neural Collapse by Design trains features and prototypes on the unit hypersphere with the simplex ETF as the global optimum of its NONL objective (Koromilas et al., 2026), our classification-only comparison baseline. On the regression side, isometric representation learning preserves data-manifold distances, including deep isomap-style extensions (Pai et al., 2022), and orthogonal label regression serves domain adaptation (Luo et al., 2024). These works constrain features, classifier weights, prototypes, or deterministic feature maps through a task loss or a fixed readout. In contrast, GPB is a conditional-prior use of distance geometry, not a new distance-preserving primitive: it declares the geometry in the mean map of the label-conditional Gaussian prior and ties the prediction readout to the same declared geometry, so the geometric code and the conditional KL are coupled during training.

Contrastive learning. Contrastive objectives organize representations by pulling same-label points together and pushing others apart, from predictive-coding formulations (van den Oord et al., 2019) and self-supervised instance discrimination (Chen et al., 2020) to supervised variants (Khosla et al., 2020), with solutions linked to neural-collapse geometry and information-bottleneck behavior (Wang & Palmer, 2024). AdaCap's permutation contrast operates at the prediction level rather than on the representation (Belucci et al., 2025), which we include as a regression-only comparison baseline. These objectives shape the representation geometry through data-driven positive–negative terms but do not include a separate criterion for controlling target-irrelevant input variation; GPB instead declares the geometry in the label-conditional prior and couples it to the KL certificate, so geometry and compression become one object rather than two.

Geometry-aware information bottlenecks. Recent work has used the phrase geometric information bottleneck in several distinct settings. A variational recommender learns a geometric prior for explanation generation (Yan et al., 2024); a data-scarcity formulation penalizes latent curvature and intrinsic dimension (Katende, 2026); and activation steering uses orthogonal subspaces to control language-model interventions (Doan et al., 2026). A concurrent analysis studies the relationship between information compression and class-wise geometric compression in CEB networks (Adilova et al., 2027). These studies make geometry an important representation concern, but they do not use the same conditional Gaussian prior for both compression and prediction, nor do they provide the OPB/EPB alignment-to- separation construction considered here.

Table 14 summarizes the distinction along these three axes: the geometric primitive is shared with several preceding lines of work, but only GPB combines a declared geometry with an information-bottleneck KL and a readout tied to the same geometry, its anchors being constrained class centers that serve simultaneously as Gaussian-prior means, KL targets, and readout anchors. The orthogonal frame is one admissible instance of the declared geometry rather than a commitment of the framework: any anchor set satisfying the two-sided separation of Eq. (5), including the centered simplex, enters the same KL objective with identical guarantees. We adopt the orthogonal convention for its stable rank-K polar construction with a full-rank identity initialization (Appendix A.3.2) and the established record of equidistant class geometries in classification (Deng et al., 2014; Xu et al., 2022).

Table 14: Comparison along the three design axes that separate GPB from its neighbors: whether the geometry is declared a priori, whether the constraint enters an information-bottleneck KL term, and whether the prediction readout is tied to the declared geometry. GPB is the only entry positive on all three; "varies" marks families that are heterogeneous on an axis.

Approach	Declared geometry	IB KL	Tied readout
Prototype methods	×	×	× (partial)
ETF / orthogonal	✓	×	✓
Fixed / hierarchy	✓	×	varies
Isometric maps	✓	×	varies
Contrastive objectives	×	×	×
IB variants (VIB, NIB, DVCCA)	×	✓	×
CEB	×	✓	×
Geometry-aware IBs	varies	✓	varies
GPB (OPB / EPB)	✓	✓	✓