# CONTRASTIVE LINEARIZED LAPLACE: LOSS GEOMETRY, VALID REPAIRS, AND RELIABLE CORES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

While contrastive learning is widely used to learn deterministic representations, it does not inherently provide estimates of uncertainty. Post-hoc uncertainty estimation methods, such as the Linearized Laplace approximation (LLA), offer a way to quantify uncertainty without retraining the encoder. However, we show that directly using the LLA induced distribution over the contrastive embedding does not yield reliable pointwise uncertainty estimates for downstream tasks: embedding variance alone does not reflect whether decisions such as nearest-neighbor retrieval or cluster assignment are stable. To address this limitation, we develop a post-hoc framework based on LLA that propagates encoder uncertainty through the downstream task. We characterize the curvature of contrastive objectives and introduce positive-semidefinite repairs where needed to obtain valid local Gaussian approximations without retraining the encoder. We test the proposed approach in metagenomic binning and image classification tasks. For metagenomic binning, posterior coassignment probabilities support consensus partitions and the selection of reliable cores, leading to improvements over existing baselines. Additional image classification experiments show calibration improvements on standard benchmarks and real histopathology datasets.

## 1 INTRODUCTION

Contrastive learning is the standard way to train representations for retrieval, verification, clustering, metric learning, and self-supervised pretraining (Hadsell et al., 2006; Chen et al., 2020a; Khosla et al., 2020; van den Oord et al., 2018). Unlike ordinary supervised prediction, these methods are trained via learning *relational* decisions: whether two examples match, which candidate is closest, or which samples belong to the same cluster. In applications such as image retrieval, genome binning, and pathology review, it is therefore important not only to obtain accurate representations, but also to identify which of these relations can be trusted. Standard contrastive encoders provide deterministic embeddings and scores, but no direct measure of how robust the resulting relational decisions are.

Bayesian inference provides a natural framework for this problem: uncertainty in the shared encoder induces joint uncertainty over all embeddings and their relations. Among approximate methods, the Linearized Laplace Approximation (LLA) is attractive because it constructs a tractable Gaussian posterior around an already trained network (Daxberger et al., 2021a; Deng et al., 2022). However, applying LLA to contrastive learning raises two difficulties. First, contrastive losses act on relational outputs (such as pair scores, embedding differences, or candidate logits) rather than on independent example-wise predictions. Hence, the resulting precision need not be positive semidefinite. Thus, a linearized encoder can yield an invalid Gaussian posterior. Second, even with a valid posterior, uncertainty in an individual embedding or pair score does not necessarily reflect uncertainty in the downstream decision, which may depend jointly on several competing relations and their posterior correlations. The key distinction is between uncertainty in a representation and uncertainty in a decision. An embedding can move substantially in the representation space (high *representation uncertainty*) while remaining grouped with the same neighbors (low *decision uncertainty*). Conversely, a small perturbation can change its assignment when it lies close to a boundary.

We address the first problem by showing that posterior validity is determined by the geometry of the contrastive loss: logit-Bernoulli and InfoNCE objectives have positive-semidefinite (PSD) output curvature, whereas distance-Bernoulli and classical margin losses can exhibit negative curvature for



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.

negative pairs. For the latter, we introduce PSD curvature constructions that make the resulting local posterior well defined. This provides a unified treatment of pairwise and candidate-set contrastive objectives and extends previous observations for Bayesian metric learning (Warburg et al., 2023).

We address the second problem by distinguishing *representation uncertainty* from *decision uncertainty*. Pointwise embedding variance is a valid posterior statistic, but it ignores decision-relevant directions and competing relations. Pairwise score variance is closer to retrieval and verification, but remains a marginal quantity. Decisions such as nearest-neighbor retrieval, graph construction, and clustering depend jointly on several scores, their margins, and their covariance. We therefore propagate representation uncertainty through the downstream rule and measure the stability of the decision uncertainty. This yields *reliable cores*: subsets of a downstream structure whose assignments remain stable under posterior perturbations, summarized through quality–coverage tradeoffs.

Our contributions are as follows. First, we formulate a relational LLA for pairwise and candidate-set contrastive losses and characterize posterior validity through their output-space geometry. Second, we provide PSD constructions for distance-based objectives with indefinite curvature. Third, we formalize why pointwise embedding variance and marginal pair-score variance are insufficient for general relational decisions. Finally, we propose posterior-sampled reliable cores that quantify uncertainty after applying the downstream rule rather than through a surrogate variance threshold. We demonstrate state-of-the-art performance on DNABERT-S-style genome clustering (Zhou et al., 2024) benchmarks. We further perform experiments in image classification on EMNIST (Cohen et al., 2017), CIFAR-100 (Krizhevsky, 2009) and SICAPv2 pathology dataset (Silva-Rodríguez, 2020).

## 2 Related Work

Uncertainty in metric and contrastive learning is often modeled through distribution-valued embeddings, including instance and face embeddings, triplet learning, cross-modal retrieval, video representations, and InfoNCE-based objectives (Oh et al., 2019; Shi and Jain, 2019; Warburg et al., 2021; Chun et al., 2021; Park et al., 2022; Kirchhof et al., 2023). Other work estimates representation reliability or uses uncertainty to select and reweight contrastive pairs during training (Park et al., 2024; Wang et al., 2023; Devalraju and Sekhar, 2025; Chang et al., 2025). These methods primarily attach uncertainty to individual representations or use it as a training signal. We instead infer an epistemic posterior over the shared encoder, inducing joint uncertainty over its relational outputs.

Most closely related, Warburg et al. (2023) applies Laplace inference to pairwise metric learning and addresses invalid local curvature, while Cheng et al. (2025) uses weight-space Laplace inference for cross-modal retrieval. We generalize this perspective to a relational score map covering both pairwise and candidate-set objectives and characterize posterior validity through output-space curvature. At the downstream level, our approach is related to posterior-sampled retrieval uncertainty (Gomez, 2025), risk-controlled retrieval (Cai et al., 2025), selective prediction (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017), and selective clustering (Marandon et al., 2025). Rather than thresholding an isolated representation or score, we propagate posterior encoder samples through the downstream relational rule and measure the stability of the resulting structure.

General Laplace and LLA methods are reviewed in Appendix A.

## 3 Laplace Approximations and Contrastive Objectives

Contrastive learning learns representations by comparing examples rather than predicting an individual target for each input. Given an encoder $h_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^d$, an input $\mathbf{x}_i$ is mapped to an embedding $\mathbf{z}_i = h_{\boldsymbol{\theta}}(\mathbf{x}_i)$. Training encourages embeddings of *positive* examples to be similar and those of *negative* examples to be dissimilar, where positive and negative relations may arise from labels, augmentations, or other notions of semantic similarity.

We denote by $t$ a contrastive training unit containing the examples and relations required to compute one loss term. Associated with $t$ is a score map $f_t(\boldsymbol{\theta}) \in \mathbb{R}^{K_t}$ and a loss $\ell_t : \mathbb{R}^{K_t} \to \mathbb{R}$, giving the training objective $\sum_t \ell_t(f_t(\boldsymbol{\theta}))$. For example, in InfoNCE (van den Oord et al., 2018), $t$ contains an anchor $i$, a candidate set $C_i$, and a positive candidate $p(i)$. Its scores are the anchor–candidate logits

$f_{t,j} = s(\mathbf{z}_i, \mathbf{z}_j)/\tau$, $j \in C_i$, and

$$\ell_t^{\text{InfoNCE}}(f_t) = -f_{t,p(i)} + \log \sum_{j \in C_i} \exp(f_{t,j}), \tag{1}$$

where $s$ is a similarity function and $\tau > 0$ is the temperature.

**Contrastive Linearized Laplace Approximation** Consider the score map $f_t(\boldsymbol{\theta})$ with loss $\ell_t(f_t(\boldsymbol{\theta}))$. Under a Gaussian parameter prior $\mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Lambda}_0^{-1})$, the MAP estimate $\boldsymbol{\theta}^\star$ minimizes

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_t \ell_t(f_t(\boldsymbol{\theta})) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{\Lambda}_0 (\boldsymbol{\theta} - \boldsymbol{\theta}_0). \tag{2}$$

The Laplace approximation constructs $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma})$ using the local curvature of this objective. For each loss term,

$$\nabla^2_{\boldsymbol{\theta}} \ell_t = \mathbf{J}_t^\top \mathbf{H}_t \mathbf{J}_t + \sum_{k=1}^{K_t} [\nabla_{f_t} \ell_t]_k \nabla^2_{\boldsymbol{\theta}} f_{t,k}, \tag{3}$$

where $\mathbf{J}_t = \partial f_t / \partial \boldsymbol{\theta}$ and $\mathbf{H}_t = \nabla^2_{f_t} \ell_t$, evaluated at $\boldsymbol{\theta}^\star$. The Linearized Laplace Approximation (LLA) replaces the model locally by $f_t(\boldsymbol{\theta}) \approx f_t(\boldsymbol{\theta}^\star) + \mathbf{J}_t(\boldsymbol{\theta} - \boldsymbol{\theta}^\star)$. The second term in Equation 3 therefore vanishes, yielding the Generalized Gauss–Newton (GGN) precision

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Lambda}_0 + \sum_t \mathbf{J}_t^\top \mathbf{H}_t \mathbf{J}_t. \tag{4}$$

**Validity of the defined precision** When $\mathbf{H}_t \succeq 0$, each data contribution is positive semidefinite and Equation 4 defines a valid Gaussian precision given a positive definite prior or damping. Thus, posterior validity depends directly on the curvature of the contrastive loss with respect to its relational scores. If $\boldsymbol{\Lambda}_0 \succeq 0$ and $\mathbf{H}_t \succeq 0$ for every contrastive unit $t$, then $\boldsymbol{\Sigma}^{-1} \succeq 0$. If, additionally, $\boldsymbol{\Lambda}_0 \succ 0$ or positive damping is added, then $\boldsymbol{\Sigma}^{-1} \succ 0$ and defines a proper local Gaussian posterior (Daxberger et al., 2021a).

For InfoNCE, choosing $f_t(\boldsymbol{\theta})$ as the vector of anchor–candidate logits satisfies this condition. This can be proved by simply writing $\mathbf{p}_t = \text{softmax}(f_t(\boldsymbol{\theta}))$,

$$\mathbf{H}_t = \text{diag}(\mathbf{p}_t) - \mathbf{p}_t \mathbf{p}_t^\top \succeq 0, \tag{5}$$

which is the covariance matrix of a categorical distribution over the candidate set. Hence InfoNCE yields a PSD GGN contribution while retaining the local coupling between candidates.

Positive semidefiniteness is not guaranteed for other contrastive objectives. Consider the pairwise margin loss (Hadsell et al., 2006)

$$f_t(\boldsymbol{\theta}) = \mathbf{z}_i - \mathbf{z}_j, \qquad \ell_t^{\text{mrg}}(f_t(\boldsymbol{\theta}), y_t) = y_t \|f_t(\boldsymbol{\theta})\|_2^2 + (1 - y_t)[m - \|f_t(\boldsymbol{\theta})\|_2]_+^2, \tag{6}$$

where $y_t = 1$ denotes a positive pair and $y_t = 0$ a negative pair. For a negative pair within the margin, $\mathbf{H}_t$ is indefinite due to negative tangential curvature; see Appendix B for further details. More generally, distance- and margin-based contrastive losses need not be convex in their relational scores. For these cases, we consider PSD repairs that either remove negative local curvature (spectral repair) or retain the informative radial geometry of distance-based losses (radial repair); their precise construction is given in Appendix B. We additionally study the Empirical Fisher as a PSD alternative in Appendix C. Unless stated otherwise, our experiments use the *spectral* repair of the GGN approximation, with all alternatives compared in Appendix G.

## 4 Marginal Variance Is Not Downstream Uncertainty

A repaired LLA defines a Gaussian posterior over the encoder parameters, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}_\theta)$. Contrastive LLA defines this posterior by locally linearizing the relational score map with respect to the encoder parameters. The resulting parameter posterior can in turn be propagated to the representation space to characterize uncertainty in individual embeddings. Specifically, for an input $\mathbf{x}$, we linearize the embedding network around the MAP parameters,

$$h_{\boldsymbol{\theta}}(\mathbf{x}) \approx h_{\boldsymbol{\theta}^\star}(\mathbf{x}) + \mathbf{J}_{\mathbf{x}}(\boldsymbol{\theta} - \boldsymbol{\theta}^\star),$$

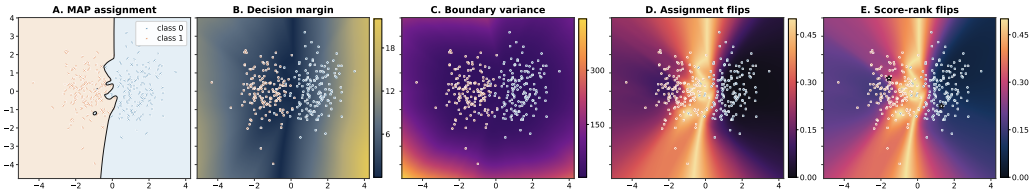



Figure 1: Synthetic two-blob example showing that posterior variance is not itself downstream uncertainty. Assignment uncertainty depends on the decision margin and variance along the boundary normal, while score-rank uncertainty depends on the posterior variance of a score difference.

which induces a Gaussian distribution in embedding space,

$$h_{\boldsymbol{\theta}}(\mathbf{x}) \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}), \qquad \boldsymbol{\mu}_{\mathbf{x}} = h_{\boldsymbol{\theta}^\star}(\mathbf{x}), \qquad \boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{J}_{\mathbf{x}} \boldsymbol{\Sigma}_{\theta} \mathbf{J}_{\mathbf{x}}^{\top}.$$

We refer to these induced distributions as Gaussian embeddings. These moments are useful, but they are not automatically uncertainties for a downstream task. A downstream task does not usually ask how much an embedding or a score varies in isolation. It asks whether a decision changes: whether an item changes class, crosses a cluster boundary, enters a graph, or changes its rank relative to another item. The relevant uncertainty is therefore tied to the margin or decision rule used by the task.

In the following theoretical results, we use a simple linear binary assignment problem to illustrate why Gaussian embedding uncertainty may not directly reflect uncertainty in a downstream task. This covers a linear classifier, and it also describes the boundary between two Euclidean nearest-centroid regions, where the sign of an affine function determines which centroid is closer. Figure 1 serves as illustrative guidance on these following results.

**Proposition 1.** *Let $h_{\boldsymbol{\theta}}$ be a (linearized) contrastive encoder and $q(\boldsymbol{\theta})$ be a (repaired) LLA posterior, then, for any $\mathbf{x} \in \mathcal{X}$, $h_{\boldsymbol{\theta}}(\mathbf{x}) \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Consider a linear binary decision boundary defined as $D(\mathbf{z}) := \mathbf{1}\{\mathbf{a}^\top \mathbf{z} + b \geq 0\}$. Then, denoting the standard gaussian c.d.f. as $\Phi$, the posterior probability of changing the assignment made at the posterior mean is*

$$\mathbb{P}[D(h(\mathbf{x})) \neq D(\boldsymbol{\mu}) \mid \mathcal{D}] = \Phi\left(-\frac{|\mathbf{a}^\top \boldsymbol{\mu} + b|}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}}}\right).$$

This result provides some intuition about the relation between Gaussian embeddings and downstream uncertainty. First, two Gaussian embeddings with the same pointwise covariance can have different assignment uncertainty if their means have different margins. Conversely, two Bayesian embeddings whose covariance matrices have the same trace can have different assignment uncertainty if their posterior variance is oriented differently relative to the boundary. The pointwise covariance is useful only after the downstream assignment has specified which direction and which margin matter. We explain this visually in Figure 4 in Appendix D.

Second, the closed-form calculation also indicates another limitation of using a single pointwise variance score in more general downstream tasks: in realistic scenarios, the effective boundary may be implicit, discrete, or defined jointly by many examples. There may be no single fixed hyperplane $(\mathbf{a}, b)$ for which the above formula is available. In such cases, the uncertainty object must be defined at the level of the downstream decision itself rather than by a universal scalar summary of $\boldsymbol{\Sigma}$.

A similar issue arises in pairwise scores. Consider the case of a retrieval task, denoting $\mathbf{x}_i$ to a query and $\mathbf{x}_j$ and $\mathbf{x}_k$ two candidate items, with the system selecting the candidate with the larger similarity score. Generally, the downstream decision depends on the comparison margin between their scores.

**Proposition 2.** *Suppose a (repaired) LLA posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma})$, the two posterior (linearized) scores are jointly Gaussian under*

$$\begin{pmatrix} s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) \\ s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_k) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{ij} \\ \mu_{ik} \end{pmatrix}, \begin{pmatrix} V_{ij} & C_{ij,ik} \\ C_{ik,ij} & V_{ik} \end{pmatrix}\right).$$

*For the downstream comparison $A_i = \mathbf{1}\{s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) \geq s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_k)\}$, the probability of changing the score rank w.r.t the MAP is given by*

$$\mathbb{P}\left[A_i(\boldsymbol{\theta}) \neq A_i(\boldsymbol{\theta}^\star)\right] = \Phi\left(-\frac{|\mu_{ij} - \mu_{ik}|}{\sqrt{V_{ij} + V_{ik} - 2C_{ij,ik}}}\right),$$

The score-level result has the same interpretation as the embedding-level result. The marginal variances $V_{ij}$ and $V_{ik}$ describe the uncertainty of two scores separately, but the comparison is determined by their difference. Hence the covariance $C_{ij,ik}$ is part of the downstream uncertainty: two uncertain scores may move together and leave their difference stable, or move against each other and make the comparison unstable. The correct variance for the ranking decision is the variance of the comparison margin, not the marginal variance of either score alone. See Figure 5 in Appendix D.

These two cases show the same principle at two levels. Pointwise embedding covariance becomes task-relevant only after it is applied to an assignment margin. Pairwise score variance becomes task-relevant only after it is applied to a score comparison or threshold margin. When these margins are explicit, the resulting stability probabilities can be written in closed form. When the downstream rule is more complex and its boundary is not available analytically, uncertainty must instead be defined by the induced posterior over the downstream decisions themselves.

## 5 Posterior Expected Downstream Tasks and Reliable Cores

The previous section shows that marginal embedding variances and marginal score variances are not, by themselves, downstream uncertainties. What we want to measure is the downstream task predictive uncertainty after encoder uncertainty has been propagated through the downstream rule. In other words, we want an expectation under the (repaired) LLA posterior over encoders. We now describe our proposal for propagating the encoder uncertainty to the downstream task.

**Classification** For classification, propagating encoder uncertainty means accounting for how different posterior encoder realizations affect the predicted class probabilities. Each draw $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$ produces an embedding of $\mathbf{x}_i$, which a downstream classifier maps to a probability vector over the classes. We denote the probability assigned to class $y$ under this realization by $p_{\boldsymbol{\theta}}(y|\mathbf{x}_i)$.

Although these probabilities may vary across posterior samples, their coordinates always refer to the same classes. We can therefore combine the predictions by averaging the probability assigned to each class separately. This gives the posterior-averaged classifier

$$\bar{p}_i(y) = \mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta})}\left[p_{\boldsymbol{\theta}}(y \mid \mathbf{x}_i)\right] \approx \tfrac{1}{S}\sum_{s=1}^{S} p_{\boldsymbol{\theta}^{(s)}}(y \mid \mathbf{x}_i) \quad , \quad \boldsymbol{\theta}^{(s)} \sim q(\boldsymbol{\theta}). \tag{7}$$

Thus, $\bar{p}_i(y)$ summarizes support for class $y$ across the encoder realizations represented by the posterior. If these realizations confidently agree on a class, the averaged prediction remains concentrated on that class; disagreement distributes probability across competing classes. When the expectation is not available analytically, it can be approximated using posterior samples.

Thus, for probabilistic classification, the downstream expectation is Bayesian model averaging of the class probabilities under the LLA posterior over encoders. Typically, that expectation cannot be computed in closed form; however, in our experiments, a logistic regressor classifier is used over the embeddings, which induces Gaussian logits. The resulting class probabilities must still be averaged over the posterior; we approximate this expectation using the Monte Carlo estimator in Equation 7.

**Clustering** The same idea cannot be applied directly to the cluster labels. Cluster labels are arbitrary: cluster 1 in one posterior sample does not necessarily correspond to cluster 1 in another posterior sample. Therefore, averaging cluster labels, or averaging one-hot encodings of cluster labels, is not meaningful without an additional label-alignment step.

Instead, we average a label-invariant clustering quantity: whether two points are assigned to the same cluster. For each posterior sample $\boldsymbol{\theta}^{(s)}$, we embed the evaluation set and apply the same clustering rule used at test time. This gives sampled assignments $\hat{y}_i^{(s)},\ i = 1, \ldots, N,\ s = 1, \ldots, S$. Although the numerical labels $\hat{y}_i^{(s)}$ are not comparable across samples, the event $\hat{y}_i^{(s)} = \hat{y}_j^{(s)}$ is comparable. We therefore build the posterior similarity matrix (PSM) (Binder, 1978), defined as

$$P_{ij} = \mathbb{P}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta})}\left[i \text{ and } j \text{ are assigned to the same cluster}\right] \approx \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\{\hat{y}_i^{(s)} = \hat{y}_j^{(s)}\}. \tag{8}$$

In other words, $P_{ij}$ is the posterior probability that points $i$ and $j$ remain together under encoder perturbations, and the clustering analogue of the posterior-averaged class-probability vector in

classification. Pairwise coassignment probabilities need not define a globally consistent partition. However, a final clustering step (e.g. spectral clustering) can be easily approached using $1 - P$ as distance function, producing final consensus labels.

### 5.1 Measuring reliability through selective retention

The posterior-averaged downstream output also provides a reliability score for selective retention. The goal is to test whether the downstream result improves after rejecting samples whose posterior behaviour is unstable. This gives a quality–coverage tradeoff: coverage is the fraction of retained inputs, and quality is measured by the downstream metric on the retained set.

For clustering, reliability is computed from the PSM in equation 8, since raw cluster labels cannot be averaged. Let $c_i$ be the consensus label obtained from the final clustering step of item $i$. We define

$$\rho_i = \frac{1}{N-1} \sum_{j \neq i} P_{ij} \, \mathbf{1}\{c_i = c_j\}. \tag{9}$$

as a measure of the robustness of the consensus assignment for item $i$. This is high when the same-cluster relations around $i$ are repeatedly recovered across posterior samples. A *reliable core* is obtained by retaining the highest-$\rho_i$ items per cluster. As a result, uncertain or border items are discarded first, while items whose cluster membership is consistently supported by the posterior are retained. Varying the retained fraction gives a quality–coverage frontier, showing the downstream result improvement when posterior-unstable inputs are discarded. This is especially useful in clustering, where unstable boundary points can obscure the reliable core structure of the partition.

**Scalability** The described approach requires computing the probability matrix of the whole inputs, which requires $\mathcal{O}(N^2)$ space. This is clearly prohibitive in real-world clustering scenarios. To alleviate this scalability issue, we approximate those quantities by a $k$-NN approach, where only the nearest $k$ inputs (in embedding space) are used to compute these quantities.

## 6 Experiments

The experiments evaluate whether the geometric construction above yields useful uncertainty estimates after the contrastive posterior is propagated through downstream decisions. Appendix E includes further details in datasets, posterior, baselines, and metrics.

The experiments follow a transfer-style, post-hoc uncertainty-estimation protocol. Rather than retraining the encoder under a Bayesian objective, we take an already trained deterministic checkpoint as the MAP solution, fit a last-layer LLA using the contrastive objective, and propagate the resulting posterior through the downstream task, this is closely related to post-hoc Laplace inference for contrastive metric learning (Warburg et al., 2023). This follows the standard use of Laplace and LLA to equip pretrained networks with uncertainty without modifying their original training procedure, often restricting inference to the last layer or a subnetwork for scalability (Daxberger et al., 2021a;b; Deng et al., 2022; Ortega et al., 2024). When the checkpoint was obtained using a different pretraining objective, the resulting posterior should therefore be interpreted as a task-adapted local posterior around the transferred representation, rather than as a posterior over the original process.

**The effect of the curvature repairs** In Table 1, we show the impact of the different repair modes proposed in Appendix B. For Bernoulli and Margin losses when applied to the CLIP encoder on both EMNIST and CIFAR100 in the classification task. The table shows the obtained proportion of negative eigenvalues, their minimum values before repair, and the change in matrix norm after repair. Spectral repair achieves substantially smaller curvature corrections than radial repair while preserving comparable downstream results. This advantage holds for both Bernoulli-distance and margin losses. See the implementation details of this experiment in Appendix E.3.

### 6.1 Metagenomic Binning

The process of organizing DNA fragments from a mixed microbial sample to recover the full (source) genomes is known as metagenomic binning (Kunin et al., 2008). SemiBin2 (Pan et al., 2023) performs this task by clustering a contrastively learned embedding space in which contigs expected

Table 1: Curvature diagnostics and downstream posterior-consensus gains (w.r.t. MAP) on EMNIST and CIFAR100 using a pretrained CLIP encoder, averaged over 5 seeds. Output-Hessian columns report the minimum eigenvalue before repair ($\lambda_{\min}$), the proportion of negative eigenvalues before repair, and the mean per-output-Hessian Frobenius change induced by the repair ($\overline{\|\Delta H\|_F}$).

| | | | Output-Hessian repair | | | Posterior change | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Loss | Repair | $\lambda_{\min}$ | Neg. rate | $\overline{\|\Delta H\|_F}$ | $\Delta$ACC ↑ | $\Delta$ECE ↓ | $\Delta$NLL ↓ |
| EMNIST | Bernoulli distance | Radial | $-6.22 \times 10^{-3}$ | 50.2% | 22.48 | +0.009 | −0.030 | −0.108 |
| | Bernoulli distance | Spectral | $-6.22 \times 10^{-3}$ | 50.2% | $3.68 \times 10^{-4}$ | +0.008 | −0.030 | −0.108 |
| | Margin | Radial | $-1.22 \times 10^{0}$ | 13.3% | 23.20 | +0.009 | −0.030 | −0.109 |
| | Margin | Spectral | $-1.22 \times 10^{0}$ | 13.3% | 0.72 | +0.009 | −0.030 | −0.108 |
| CIFAR-100 | Bernoulli distance | Radial | $-3.32 \times 10^{-11}$ | 49.3% | 22.60 | +0.021 | −0.080 | −0.451 |
| | Bernoulli distance | Spectral | $-3.32 \times 10^{-11}$ | 49.3% | $9.13 \times 10^{-13}$ | +0.021 | −0.080 | −0.452 |
| | Margin | Radial | $-9.92 \times 10^{-1}$ | 14.0% | 23.24 | +0.022 | −0.080 | −0.450 |
| | Margin | Spectral | $-9.92 \times 10^{-1}$ | 14.0% | 0.64 | +0.022 | −0.080 | −0.449 |

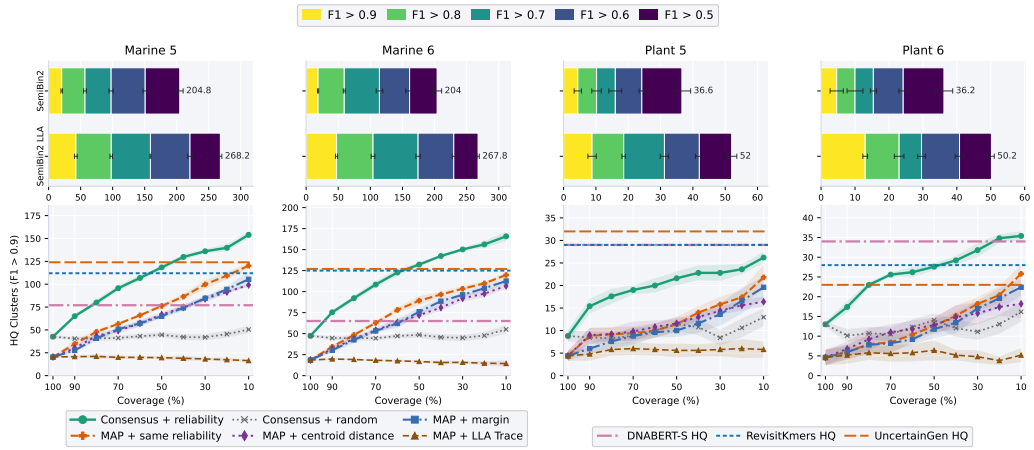



Figure 2: Genome reliable-core results on Marine and Plant splits. Top panels compare high-quality cluster counts for SemiBin2 and the LLA posterior-consensus variant across F1 thresholds. Bottom panels show high-quality clusters retained as coverage is reduced with different approaches.

to originate from the same genome are brought closer together, while contigs expected to belong to different genomes are pushed apart. This is achieved through a pairwise margin loss, shown in Equation 6, making SemiBin2 a natural setting for evaluating our relational formulation. We use this contrastive structure to study whether the proposed approach can provide meaningful uncertainty over the learned pairwise relationships and, ultimately, over the resulting metagenomic representation.

We do not train a new contrastive binning model from scratch. Instead, we use the official pretrained SemiBin2 encoder as the MAP reference model and fit a repaired (using spectral repair) last-layer KFAC LLA posterior around this checkpoint. For each posterior draw, we re-embed the contigs and apply $k$-means with $k = 500$. We estimate the posterior coassignment probability $P_{ij}$ by the fraction of sampled partitions in which contigs $i$ and $j$ are assigned to the same cluster. A consensus partition is then obtained by applying spectral clustering to the 15-nearest-neighbor graph constructed from $P$.[1] We evaluate on the Marine 5, Marine 6, Plant 5, and Plant 6 benchmark splits introduced with DNABERT-S (Zhou et al., 2024), where the quality of a cluster is measured in terms of its F1 score, naming high quality (HQ) clusters to those with $F1 > 0.9$.

Figure 2 (top) compares the SemiBin2 MAP clustering with the LLA posterior-consensus clustering across cluster-level $F_1$ thresholds. The figure shows that posterior LLA consensus (SemiBin2-LLA) increases the number of HQ clusters on all four datasets compared to the default MAP model (SemiBin2). The gains are not restricted to a single F1 threshold: the LLA variant shifts the full high-

[1] Appendix G shows the obtained results for different values of $k$.

Table 2: Classification results on CIFAR-100, EMNIST, and SICAP. Values are mean ± one standard deviation over five seeds. The best result for each dataset and encoder is shown in bold. In the second encoder block, SimCLR is used for CIFAR-100 and EMNIST, and CONCH for SICAP.

| | | CIFAR-100 | | | EMNIST | | | SICAP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Encoder | Method | Acc. ↑ | NLL ↓ | ECE ↓ | Acc. ↑ | NLL ↓ | ECE ↓ | Acc. ↑ | NLL ↓ | ECE ↓ |
| CLIP | MAP | $74.00 \pm 0.00$ | $1.424 \pm 0.000$ | $0.165 \pm 0.000$ | $84.45 \pm 0.13$ | $0.533 \pm 0.010$ | $0.064 \pm 0.001$ | $64.44 \pm 0.04$ | $1.095 \pm 0.000$ | $0.175 \pm 0.000$ |
| | LLA Sampled | $76.32 \pm 0.05$ | $0.993 \pm 0.004$ | $0.084 \pm 0.000$ | $\mathbf{85.63 \pm 0.11}$ | $0.454 \pm 0.007$ | $0.035 \pm 0.001$ | $66.70 \pm 0.45$ | $0.930 \pm 0.008$ | $0.123 \pm 0.004$ |
| | Gaussian LLA | $\mathbf{77.34 \pm 0.21}$ | $\mathbf{0.766 \pm 0.002}$ | $\mathbf{0.014 \pm 0.003}$ | $84.95 \pm 0.10$ | $\mathbf{0.449 \pm 0.002}$ | $\mathbf{0.020 \pm 0.002}$ | $\mathbf{66.76 \pm 1.51}$ | $\mathbf{0.824 \pm 0.042}$ | $\mathbf{0.062 \pm 0.018}$ |
| SimCLR / CONCH | MAP | $68.25 \pm 0.01$ | $1.774 \pm 0.012$ | $0.199 \pm 0.001$ | $\mathbf{81.79 \pm 0.29}$ | $0.686 \pm 0.017$ | $0.079 \pm 0.002$ | $79.90 \pm 0.03$ | $0.567 \pm 0.001$ | $0.093 \pm 0.001$ |
| | LLA Sampled | $\mathbf{69.65 \pm 0.12}$ | $1.451 \pm 0.004$ | $0.141 \pm 0.001$ | $81.74 \pm 0.41$ | $0.688 \pm 0.014$ | $0.080 \pm 0.004$ | $80.44 \pm 0.24$ | $0.545 \pm 0.004$ | $0.082 \pm 0.003$ |
| | Gaussian LLA | $66.74 \pm 0.20$ | $\mathbf{1.188 \pm 0.002}$ | $\mathbf{0.107 \pm 0.003}$ | $79.82 \pm 0.57$ | $\mathbf{0.673 \pm 0.047}$ | $\mathbf{0.055 \pm 0.043}$ | $\mathbf{80.88 \pm 0.15}$ | $\mathbf{0.478 \pm 0.012}$ | $\mathbf{0.045 \pm 0.018}$ |

quality-cluster profile upward relative to the MAP baseline. This indicates that posterior sampling is not merely filtering isolated ambiguous contigs, but stabilizing the recovered bin structure.

Figure 2 (bottom) evaluates selective clustering. The coverage curves show the number of HQ clusters obtained when progressively retaining only a fraction of the contigs (see Section 5.1). Here, *coverage* is the percentage of contigs retained within each predicted cluster: $100\%$ keeps all contigs, while $10\%$ keeps only the $10\%$ ranked as most reliable (cf. Equation 9). The curves compare different ways of performing this selection. We consider two variants based on consensus clustering: *Consensus + reliability* uses the consensus clustering and ranks contigs by the reliability score $\rho_i$, while *consensus + random* uses the same clustering but rejects contigs randomly. As baseline, we use the SemiBin2 MAP-based clustering combined with three different ranking strategies based on geometric distance (*MAP + centroid distance*), assignment margin (*MAP + margin*), and pointwise posterior variance (*MAP + LLA trace*). The horizontal lines report the HQ-cluster counts of the reference methods, namely DNABERT-S HQ (Zhou et al., 2024), RevisitKmers (Celikkanat et al., 2024) and UncertainGen (Celikkanat et al., 2025). Besides showing as full horizontal lines, neither of these methods present a 100% coverage, as all of them reject contigs during clustering, however, this coverage fraction is not reported in either of the works.

Across all four datasets, *consensus + reliability* produces substantially more HQ clusters as coverage decreases than random rejection or the MAP-based heuristics, and surpasses the reference methods in every dataset except for Plant 5. Using the same reliability ranking with MAP assignments also improves over most pointwise heuristics, but remains well below the posterior-consensus construction. These results support, in the real-world metagenomic binning task, the theoretical results presented in this work: marginal or geometric uncertainty scores are weak proxies for a relational decision, whereas posterior agreement after applying the clustering rule directly captures whether the resulting groups remain stable under encoder perturbations.

### 6.2 Image Classification

For the image experiments, we apply the proposed approach to pretrained visual representations and study how uncertainty in the learned embedding space propagates to downstream classification and clustering. We consider CIFAR-100 and EMNIST using CLIP (Radford et al., 2021) and SimCLR (Chen et al., 2020a) representations, together with SICAPv2 prostate histopathology patches (Silva-Rodríguez et al., 2020), where we use CLIP and CONCH-v1.5, a vision–language foundation model for computational pathology (Lu et al., 2024). The classification experiments test whether posterior uncertainty over the representation improves predictive uncertainty after a downstream readout, while clustering provides a relational setting in which reliability depends on the stability of the groups induced by the embedding space.

Again, we use each pretrained encoder as the MAP reference model and fit a GGN KFAC last-layer LLA posterior around this checkpoint (Martens and Grosse, 2015). For classification, *MAP* trains a multinomial logistic regressor on the deterministic embeddings. *Sampled LLA* samples encoders from the posterior, trains the same logistic regressor in each sampled embedding space, and averages the resulting predictive probabilities. *Gaussian LLA* places a KFAC Gaussian posterior over the parameters of the final linear embedding layer. The resulting uncertainty is propagated through the linear logistic classifier to obtain Gaussian logits. We train the classifier with Adam (Kingma and Ba, 2015) using Monte Carlo estimates of the marginal predictive likelihood.

Table 2 shows that posterior propagation improves predictive uncertainty over the MAP baseline in most cases. For CIFAR-100 and EMNIST, the LLA variants generally reduce NLL and ECE for both

CLIP and SimCLR representations, while the corresponding changes in accuracy are smaller. The same behavior is observed on SICAP: CONCH provides substantially stronger predictions than the general-purpose CLIP encoder, reflecting the benefit of pathology-specific pretraining, but posterior propagation still improves uncertainty quality for both representation families. In particular, sampled and closed-form LLA generally reduce NLL and ECE, with the latter producing the strongest overall calibration. Thus, the main effect of posterior propagation is improved predictive uncertainty rather than a large change in the MAP decisions.

## 7 Discussion, Limitations, and Conclusion

Contrastive objectives that behave similarly during MAP training can nevertheless induce different local Bayesian models. For this reason, the natural output variable of the loss, the PSD status of its output-Hessian blocks, and the precision spectrum before and after repair should be reported explicitly. These diagnostics help distinguish a valid relational LLA construction from a numerical procedure that produces a PSD matrix while obscuring the underlying loss geometry.

Our analysis is local to a trained solution, and the scalable experiments rely on last-layer or KFAC last-layer LLA rather than a full-network posterior. The proposed repairs yield valid Gaussian precisions, but they remain local approximations to objectives that are not always canonical likelihoods. They should therefore be interpreted as local modeling choices rather than as evidence that the original loss defines a well-specified likelihood. Similarly, reliable cores are task-specific summaries of uncertainty under a particular downstream rule and posterior, rather than universal uncertainty measures.

Overall, contrastive LLA raises two distinct questions. The first is geometric: does the loss provide valid local curvature? Logit-Bernoulli and InfoNCE do, whereas distance-Bernoulli and margin losses generally require repair. The second is operational: how should the resulting posterior be used downstream? Pointwise embedding variance and marginal pair-score variance are generally insufficient because they do not account for the structure of the downstream decision. Propagating the posterior through the downstream rule, for example by sampling repaired encoders and constructing reliable relational cores, instead yields uncertainty measures and quality–coverage tradeoffs that are directly tied to the decisions of interest.

**Limitations.** (i) Our posterior should be interpreted as a task-adapted contrastive LLA posterior, rather than as a full Bayesian posterior over the original pretrained model. This distinction is particularly important for experiments based on CLIP, SimCLR, DNABERT-S-style, or CONCH embeddings, since the LLA is fitted after transfer using the available downstream representation and objective, rather than the original pretraining data and loss. Consequently, the posterior captures uncertainty in the transferred embedding/readout map under the fitted last-layer or KFAC approximation, but not all sources of uncertainty in the pretrained feature extractor. (ii) The proposed spectral and radial repairs ensure a valid Gaussian precision, but they also modify the local curvature of distance- and margin-based losses. They should therefore be viewed as local modeling policies rather than as corrections to a fully specified likelihood. (iii) Reliable cores are tied to a particular downstream rule and selection procedure. They measure whether clustering, classification, or grouping decisions remain stable under posterior encoder perturbations, but they are not universal uncertainty scores, calibrated out-of-distribution detectors, or guarantees of semantic correctness. In addition, posterior propagation requires either sampling encoders or propagating Gaussian embeddings through the downstream rule, making it more expensive than computing a single marginal variance score.

### AI use statement

In this work, we used generative AI tools to provide feedback on research methodology and experimental design and to assist with the interpretation and presentation of experimental results. We did not use generative AI tools to generate datasets, formulate mathematical claims, develop theoretical models, provide critical ingredients for or assist in writing proofs, propose hypotheses, implement the proposed methods, clean or reformat datasets, or perform qualitative or thematic data analysis. Translation and research tasks involving surveys or interviews were not applicable to this work.

Additionally, we used generative AI tools for brainstorming, searching for and summarizing potentially relevant literature, suggesting improvements to the structure and presentation of the paper,

and drafting or editing text for clarity and readability. All methodological suggestions and interpretations produced with AI assistance were independently evaluated by the authors. All references were manually verified against their original sources, all experimental results and numerical claims were checked against the outputs of experiments conducted by the authors, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### Ethics statement

This work develops uncertainty estimation methods using existing benchmark datasets and pretrained encoders. The SICAPv2 experiments use publicly available histopathology data with patient-separated training and test partitions; they provide a methodological evaluation and do not establish clinical suitability. Uncertainty estimates may inherit biases from the underlying data and pretrained representations, and posterior stability does not guarantee correctness. Furthermore, selecting reliable cores may disproportionately exclude rare or underrepresented examples. Applications should therefore assess both predictive performance and the effects of selective rejection across relevant classes and populations, particularly in sensitive domains.

### Reproducibility statement

Sections 3-5 describe the relational LLA construction, posterior propagation, and reliable-core selection. Appendices B-D provide additional curvature derivations, repair definitions, and analysis of the distinction between marginal variance and downstream uncertainty. Appendix E documents the datasets, preprocessing and data splits, encoder architectures, posterior target layers, contrastive objectives, hyperparameter selection, posterior sampling, downstream rules, and evaluation metrics. Experiments use five random seeds unless otherwise stated. Appendices F-G report additional clustering results and ablations examining the sensitivity of the method to its modeling and downstream choices.

## References

Federico Bergamin, Pablo Moreno-Muñoz, Søren Hauberg, and Georgios Arvanitidis. Riemannian laplace approximations for bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 31066–31095. Curran Associates, Inc., 2023. doi: 10.52202/075280-1354. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/631f99d8e860054410c239fc90d18270-Abstract-Conference.html.

D. A. Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 04 1978. ISSN 0006-3444. doi: 10.1093/biomet/65.1.31. URL https://doi.org/10.1093/biomet/65.1.31.

Kaiwen Cai, Chris Xiaoxuan Lu, Xingyu Zhao, Wei Huang, and Xiaowei Huang. Risk controlled image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27224–27232, 2025. doi: 10.1609/aaai.v39i26.34931.

Sergio Calvo-Ordoñez, Jonathan Plenk, Richard Bergna, Álvaro Cartea, Yarin Gal, José Miguel Hernández-Lobato, and Kamil Ciosek. Richer bayesian last layers with subsampled NTK features. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2602.01279. To appear.

Abdulkadir Celikkanat, Andres R Masegosa, and Thomas D Nielsen. Revisiting k-mer profile for effective and scalable genome representation learning. *Advances in Neural Information Processing Systems*, 37:118930–118952, 2024.

Abdulkadir Celikkanat, Andres R Masegosa, Mads Albertsen, and Thomas D Nielsen. Uncertaingen: Uncertainty-aware representations of dna sequences for metagenomic binning. *arXiv preprint arXiv:2509.26116*, 2025.

Luyao Chang, Leiting Chen, and Chuan Zhou. Uncertainty-aware contrastive learning for deep clustering. *Neurocomputing*, 647:130568, 2025. doi: 10.1016/j.neucom.2025.130568.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020a. URL https://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html.

Kun Cheng, Qibing Qin, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep probabilistic binary embedding via learning reliable uncertainty for cross-modal retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6393–6402. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3754811.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966217.

Rodrigo Pereira David. Tubular riemannian laplace approximations for bayesian neural networks, 2025. URL https://arxiv.org/abs/2512.24381.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20089–20103. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/a7c9585703d275249f30a088cebba0ad-Abstract.html.

Erik Daxberger, Eric Nalisnick, James U. Allingham, Javier Antorán, and José Miguel Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2510–2521. PMLR, 2021b.

Zhijie Deng, Feng Zhou, and Jun Zhu. Accelerated linearized Laplace approximation for Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 2695–2708. Curran Associates, Inc., 2022. doi: 10.52202/068431-0195. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/12143893d9d37c3569dda800b95cabd9-Abstract-Conference.html.

Dhanunjaya Varma Devalraju and C. Chandra Sekhar. Uncertainty-guided metric learning without labels. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 7029–7038, 2025.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Lluis Gomez. Over the top-1: Uncertainty-aware cross-modal retrieval with CLIP. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1521–1532. PMLR, 2025.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006. doi: 10.1109/CVPR.2006.100.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 703–711. PMLR, 2021. URL https://proceedings.mlr.press/v130/immer21a.html.

Pedro Jiménez García-Ligero, Luis A. Ortega, Pablo Morales-Álvarez, and Daniel Hernández-Lobato. Improving the linearized Laplace approximation via quadratic approximations. In *ESANN 2026 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 721–726. Ciaco–i6doc.com, 2026. ISBN 9782875870964. doi: 10.14428/esann/2026.ES2026-154. URL https://doi.org/10.14428/esann/2026.ES2026-154.

Dongwan D Kang, Jeff Froula, Rob Egan, and Zhong Wang. Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17085–17104. PMLR, 2023.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being bayesian, even just a bit, fixes overconfidence in ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5436–5446. PMLR, 2020. URL https://proceedings.mlr.press/v119/kristiadi20a.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. M.sc. thesis, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, 72(4): 557–578, December 2008. doi: 10.1128/mmbr.00009-08.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Ariane Marandon, Tabea Rebafka, Etienne Roquain, and Nataliya Sokolovska. False selection rate control in mixture models. *Scandinavian Journal of Statistics*, 52(4):2014–2060, 2025. doi: 10.1111/sjos.70017.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417. PMLR, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Seong Joon Oh, Kevin P. Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C. Gallagher. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*, 2019.

Luis A. Ortega. *Uncertainty Estimation and Generalization Bounds for Modern Deep Learning*. PhD thesis, Universidad Autónoma de Madrid, 2026. URL https://arxiv.org/abs/2606.13818.

Luis A. Ortega, Simon Rodriguez Santana, and Daniel Hernández-Lobato. Variational linearized Laplace approximation for Bayesian deep learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38815–38836. PMLR, 2024. URL https://proceedings.mlr.press/v235/ortega24a.html.

Luis A. Ortega, Simón Rodríguez-Santana, and Daniel Hernández-Lobato. Scalable linearized Laplace approximation via surrogate neural kernel. In *ESANN 2026 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 709–714. Ciaco–i6doc.com, 2026. ISBN 9782875870964. doi: 10.14428/esann/2026.ES2026-71. URL https://doi.org/10.14428/esann/2026.ES2026-71.

Shaojun Pan, Xing-Ming Zhao, and Luis Pedro Coelho. SemiBin2: Self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics*, 39, 2023. doi: 10.1093/bioinformatics/btad209. URL https://doi.org/10.1093/bioinformatics/btad209.

Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Probabilistic representations for video contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14711–14721, 2022.

Young-Jin Park, Hao Wang, Shervin Ardeshir, and Navid Azizan. Quantifying representation reliability in self-supervised learning models. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2835–2860. PMLR, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019.

Julio Silva-Rodríguez. SICAPv2 – prostate whole slide images with gleason grades annotations. Mendeley Data, 2020. URL https://data.mendeley.com/datasets/9xxm58dvs3/1. Version 1.

Julio Silva-Rodríguez, Adrián Colomer, María A. Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195: 105637, 2020. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105637. URL https://www.sciencedirect.com/science/article/pii/S016926072031470X.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Hao Wang, Min Li, Yangyang Song, Youjian Zhang, and Liying Chi. UCoL: Unsupervised learning of discriminative facial representations via uncertainty-aware contrast. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2510–2518, 2023. doi: 10.1609/aaai.v37i2.25348.

Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12158–12168, 2021. doi: 10.1109/ICCV48922.2021.01194.

Frederik Warburg, Marco Miani, Silas Brack, and Søren Hauberg. Bayesian metric learning for uncertainty quantification in image retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pages 69178–69190. Curran Associates, Inc., 2023. doi: 10.52202/075280-3028.

Zhihan Zhou, Weimin Wu, Harrison Ho, Jiayi Wang, Lizhen Shi, Ramana V. Davuluri, Zhong Wang, and Han Liu. DNABERT-S: Pioneering species differentiation with species-aware DNA embeddings, 2024. URL https://arxiv.org/abs/2402.08777.

# A Additional Related Work

The Laplace approximation represents the posterior near a mode by a Gaussian whose precision is determined by the local curvature of the negative log posterior. In neural networks, this curvature is commonly approximated or structured using generalized Gauss–Newton (GGN), Fisher, Kronecker-factored, diagonal, low-rank, last-layer, or subnetwork approximations (Daxberger et al., 2021a; Martens, 2020; Kristiadi et al., 2020; Daxberger et al., 2021b).

A useful modern interpretation is that GGN-based Laplace performs Bayesian inference in a locally linearized generalized linear model, rather than simply approximating the Hessian of the original nonlinear network (Immer et al., 2021). This view has motivated several scalable Linearized Laplace Approximation (LLA) variants, including variational formulations, low-rank feature approximations, surrogate neural-kernel approximations, and quadratic corrections to the local posterior precision (Ortega, 2026; Deng et al., 2022; Ortega et al., 2026; Jiménez García-Ligero et al., 2026). Related work has also enriched Bayesian last layers using subsampled neural-tangent features, retaining a tractable local linear model while providing greater expressivity than standard last-layer Laplace (Calvo-Ordoñez et al., 2026).

Beyond Euclidean Gaussian approximations, Riemannian and tubular Riemannian Laplace methods adapt posterior geometry to curved, anisotropic, and symmetry-induced low-loss regions of the neural-network loss landscape (Bergamin et al., 2023; David, 2025). These approaches have primarily been developed for supervised likelihoods that decompose over individual examples. In contrast, our setting applies LLA to relational objectives, where each loss term depends jointly on pairs or sets of network outputs and the relevant curvature therefore lives in the output space of a contrastive unit.

# B Contrastive Curvature Pathologies and Repairs

This appendix develops the loss-level details of the relational LLA construction in Section 3. For a relational unit $t$, the linearized Laplace precision contains blocks of the form $\mathbf{J}_t^\top \mathbf{H}_t \mathbf{J}_t$, where $\mathbf{H}_t = \nabla^2_{f_t} \ell_t(f_t)$ and $f_t$ is the output variable used by the contrastive loss. The choice of this variable matters. Logit losses define curvature with respect to a scalar score or a vector of candidate logits, whereas distance losses define it with respect to an embedding difference. Although these choices can have little visible effect during training, they can lead to very different local Bayesian models.

Throughout this section, we write $\mathbf{z}_i = h_{\boldsymbol{\theta}}(\mathbf{x}_i)$, $s_{ij} = \tau^{-1} \mathbf{z}_i^\top \mathbf{z}_j$, and $\mathbf{v}_{ij} = \mathbf{z}_i - \mathbf{z}_j$, where $\tau > 0$ is the temperature. All statements about curvature refer to the output space, before adding the prior, damping, or numerical jitter. We use $y = 1$ for a positive pair and $y = 0$ for a negative pair.

## B.1 Logit losses give PSD output curvature

**Logit-Bernoulli.** For a scalar pair score $s_{ij}$, the Bernoulli negative log-likelihood is $\ell^{\text{logit}}(s_{ij}, y) = \text{softplus}(s_{ij}) - y s_{ij}$. Writing $p = \sigma(s_{ij})$, its first derivative is $\partial \ell^{\text{logit}} / \partial s_{ij} = p - y$, and its second derivative is $\partial^2 \ell^{\text{logit}} / \partial s_{ij}^2 = p(1-p) \geq 0$. Each output curvature block is therefore PSD, as is the corresponding GGN contribution $p(1-p)\mathbf{J}_t^\top \mathbf{J}_t$, regardless of the Jacobian $\mathbf{J}_t$.

**InfoNCE.** For an anchor $i$ with candidate set $C_i$ and positive candidate $p(i) \in C_i$, the output is the vector of candidate logits $f_t = (a_{t,j})_{j \in C_i}$, with $a_{t,j} = s_{ij}$. The InfoNCE loss is

$$\ell^{\text{InfoNCE}}(f_t) = -a_{t,p(i)} + \log \sum_{j \in C_i} \exp(a_{t,j}).$$

If $\mathbf{q}_i = \text{softmax}(f_t)$, the gradient is $\nabla_{f_t} \ell^{\text{InfoNCE}} = \mathbf{q}_i - \mathbf{e}_{p(i)}$, and the Hessian is

$$\nabla^2_{f_t} \ell^{\text{InfoNCE}} = \text{diag}(\mathbf{q}_i) - \mathbf{q}_i \mathbf{q}_i^\top.$$

This is the covariance matrix of the one-hot representation of a categorical variable with probabilities $\mathbf{q}_i$. Equivalently, for any vector $\mathbf{r}$,

$$\mathbf{r}^\top \big(\text{diag}(\mathbf{q}_i) - \mathbf{q}_i \mathbf{q}_i^\top\big)\mathbf{r} = \sum_{j \in C_i} q_{i,j} r_j^2 - \left( \sum_{j \in C_i} q_{i,j} r_j \right)^2 = \text{Var}_{j \sim \mathbf{q}_i}(r_j) \geq 0.$$

InfoNCE therefore also gives a PSD output curvature block. Its rank is at most $|C_i| - 1$, since adding the same constant to every logit leaves the softmax unchanged. This rank deficiency is not a pathology: the prior and numerical jitter make the final posterior precision invertible.

### B.2 Distance losses can give indefinite output curvature

For distance-based losses, the output variable is the embedding displacement $\mathbf{v} = \mathbf{v}_{ij} = \mathbf{z}_i - \mathbf{z}_j$. We therefore need the curvature with respect to $\mathbf{v}$, rather than a scalar logit, and it is in this space that negative curvature can arise.

To see where it comes from, consider a radial loss $\ell(\mathbf{v}) = g(r)$, where $r = \|\mathbf{v}\|_2 > 0$ and $\mathbf{u} = \mathbf{v}/r$. Its Hessian is

$$\nabla^2_{\mathbf{v}} \ell(\mathbf{v}) = g''(r)\mathbf{u}\mathbf{u}^\top + \frac{g'(r)}{r}\big(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top\big).$$

There is one radial eigenvalue, along $\mathbf{u}$, and $d-1$ tangential eigenvalues, in directions orthogonal to $\mathbf{u}$. Tangential perturbations change the orientation of the difference vector to first order. A negative tangential eigenvalue means that the loss is locally concave along such a direction, so a small straight-line displacement in that direction decreases the loss. This negative curvature cannot be used as Gaussian precision.

**Distance-Bernoulli.** The distance-Bernoulli model sets $p(\mathbf{v}) = \exp(-\|\mathbf{v}\|_2^2)$. For a positive pair ($y = 1$), the loss is $\ell_+^{\mathrm{dist}}(\mathbf{v}) = -\log p(\mathbf{v}) = \|\mathbf{v}\|_2^2$, with Hessian $\nabla^2_{\mathbf{v}} \ell_+^{\mathrm{dist}}(\mathbf{v}) = 2\mathbf{I}_d \succeq 0$. Positive pairs therefore do not cause a problem with positive definiteness.

For a negative pair ($y = 0$), the loss becomes

$$\ell_-^{\mathrm{dist}}(\mathbf{v}) = -\log\big(1 - \exp(-\|\mathbf{v}\|_2^2)\big).$$

Writing $p = \exp(-\|\mathbf{v}\|_2^2)$ and $q = 1 - p$, its Hessian for $\mathbf{v} \neq \mathbf{0}$ is

$$\mathbf{H}_-^{\mathrm{dist}}(\mathbf{v}) = \nabla^2_{\mathbf{v}} \ell_-^{\mathrm{dist}}(\mathbf{v}) = 4\frac{p}{q^2}\mathbf{v}\mathbf{v}^\top - 2\frac{p}{q}\mathbf{I}_d. \tag{10}$$

The radial eigenvalue is

$$\lambda_{\mathrm{dist,rad}} = 4\frac{p}{q^2}\|\mathbf{v}\|_2^2 - 2\frac{p}{q} = 2\frac{p}{q^2}\big(2\|\mathbf{v}\|_2^2 - q\big),$$

which is positive whenever $\mathbf{v} \neq \mathbf{0}$. Every tangential eigenvalue, however, is $\lambda_{\mathrm{dist,tan}} = -2p/q < 0$. The output Hessian of a negative distance-Bernoulli pair is therefore indefinite when the embedding dimension is $d \geq 2$.

**Pairwise margin loss.** The standard pairwise margin loss is

$$\ell^{\mathrm{mrg}}(\mathbf{v}, y) = y\|\mathbf{v}\|_2^2 + (1-y)[m - \|\mathbf{v}\|_2]_+^2.$$

For positive pairs, we again have $\ell_+^{\mathrm{mrg}}(\mathbf{v}) = \|\mathbf{v}\|_2^2$ and $\nabla^2_{\mathbf{v}} \ell_+^{\mathrm{mrg}}(\mathbf{v}) = 2\mathbf{I}_d$, so the Hessian is PSD. Inactive negatives, with $\|\mathbf{v}\|_2 \geq m$, are assigned zero curvature, with the boundary case discussed below. The problematic case is an active negative with $0 < r = \|\mathbf{v}\|_2 < m$. Its loss is $\ell_-^{\mathrm{mrg}}(\mathbf{v}) = (m - r)^2$, and its Hessian is

$$\mathbf{H}_-^{\mathrm{mrg}}(\mathbf{v}) = 2\mathbf{u}\mathbf{u}^\top + \frac{2(r-m)}{r}\big(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top\big), \qquad \mathbf{u} = \mathbf{v}/r. \tag{11}$$

The radial eigenvalue is 2, but every tangential eigenvalue is $\lambda_{\mathrm{mrg,tan}} = 2(r-m)/r < 0$. Active negative margin pairs thus also have indefinite Hessians whenever $d \geq 2$.

At $r = 0$ and $r = m$, the classical Hessian is not defined. In practice, these points are either avoided almost surely or treated using standard numerical stabilizers, such as replacing $r$ by $\sqrt{\|\mathbf{v}\|_2^2 + \epsilon}$ in diagnostic computations. The negative curvature does not arise from these exceptional points: it is present throughout the active region $0 < r < m$.

**Proposition 3.** *Let $\mathbf{H}_t$ be the output Hessian of a training unit and $\mathbf{J}_t$ its last-layer Jacobian. If a parameter perturbation $\mathbf{u}$ satisfies $(\mathbf{J}_t\mathbf{u})^\top \mathbf{H}_t (\mathbf{J}_t\mathbf{u}) < 0$, then $\mathbf{J}_t^\top \mathbf{H}_t \mathbf{J}_t$ is not positive semidefinite. In particular, consider a negative distance-Bernoulli pair with $\mathbf{v} \neq \mathbf{0}$, or an active margin negative with $0 < \|\mathbf{v}\|_2 < m$. When $d \geq 2$, any nonzero tangential direction $\mathbf{w} \perp \mathbf{v}$ that can be written as $\mathbf{w} = \mathbf{J}_t\mathbf{u}$ gives negative curvature.*

*Proof.* For any parameter perturbation $\mathbf{u}$,

$$\mathbf{u}^\top \mathbf{J}_t^\top \mathbf{H}_t \mathbf{J}_t \mathbf{u} = (\mathbf{J}_t \mathbf{u})^\top \mathbf{H}_t (\mathbf{J}_t \mathbf{u}).$$

If the right-hand side is negative, the parameter-space block has a negative quadratic form in the direction $\mathbf{u}$ and therefore cannot be PSD. □

Linearizing the encoder therefore does not, by itself, resolve the problem with distance losses. LLA removes second derivatives of the network or relational score map, but still uses the output Hessian of the loss. When this Hessian is indefinite, the corresponding GGN block can be indefinite too. Contributions from other training units, the prior, or sufficiently large damping may hide this in the assembled precision, but the loss itself no longer guarantees a valid Gaussian precision.

### B.3 PSD OUTPUT-CURVATURE REPAIRS

Because the negative curvature arises within the output Hessian of a contrastive unit, we repair each block before accumulating it in parameter space. We keep the MAP encoder and relational Jacobian fixed and replace only $\mathbf{H}_t$ by a PSD block $\widetilde{\mathbf{H}}_t \succeq 0$. This gives the repaired precision

$$\widetilde{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Lambda}_0 + \sum_t \mathbf{J}_t^\top \widetilde{\mathbf{H}}_t \mathbf{J}_t + \epsilon \mathbf{I}, \tag{12}$$

where $\boldsymbol{\Lambda}_0$ includes the prior or weight-decay precision and $\epsilon > 0$ is a small numerical jitter. Under the default spectral repair described below, blocks that are already PSD, including those from logit-Bernoulli and InfoNCE, are left unchanged up to numerical tolerance.

A repair makes the local Gaussian precision valid, but its interpretation remains local. It does not retrain the encoder or change the MAP embeddings, nor does it establish that the original distance or margin loss was a fully specified likelihood. It is a modeling choice that turns an indefinite contrastive curvature block into a usable PSD precision block. We consider two such repairs.

**Spectral repair.** Our default repair is the spectral projection onto the PSD cone, which can be applied regardless of the loss. We first remove any numerical asymmetry by setting $\mathbf{H}_{t,\mathrm{sym}} = \frac{1}{2}(\mathbf{H}_t + \mathbf{H}_t^\top)$. If its eigendecomposition is $\mathbf{H}_{t,\mathrm{sym}} = \mathbf{Q}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_m)\mathbf{Q}^\top$, the repaired block is

$$\widetilde{\mathbf{H}}_t^{\mathrm{spec}} = \Pi_+(\mathbf{H}_t) = \mathbf{Q}\,\mathrm{diag}([\lambda_1]_+, \ldots, [\lambda_m]_+)\mathbf{Q}^\top, \qquad [a]_+ = \max\{a, 0\}. \tag{13}$$

This retains every nonnegative curvature direction supplied by the loss and removes the negative directions that could make the Gaussian precision indefinite. It differs from the absolute-value transform $\lambda_i \mapsto |\lambda_i|$: negative curvature is discarded, rather than converted into positive precision. The repair can therefore reduce the rank of a block. The prior and jitter in equation 12 make the final precision invertible.

**Radial-only repair for distance-based losses.** For a distance-based loss, the output is the embedding difference $\mathbf{v}$, and its eigenspaces are available analytically: the radial direction $\mathbf{u} = \mathbf{v}/\|\mathbf{v}\|_2$ and the $d-1$ tangential directions orthogonal to it. Let $\mathbf{P}_{\mathrm{rad}} = \mathbf{u}\mathbf{u}^\top$ and $\mathbf{P}_{\mathrm{tan}} = \mathbf{I}_d - \mathbf{u}\mathbf{u}^\top$. For a radial Hessian $\mathbf{H}(\mathbf{v}) = \lambda_{\mathrm{rad}}\mathbf{P}_{\mathrm{rad}} + \lambda_{\mathrm{tan}}\mathbf{P}_{\mathrm{tan}}$, spectral repair retains all nonnegative eigenvalues:

$$\widetilde{\mathbf{H}}_{\mathrm{spec}}(\mathbf{v}) = [\lambda_{\mathrm{rad}}]_+\mathbf{P}_{\mathrm{rad}} + [\lambda_{\mathrm{tan}}]_+\mathbf{P}_{\mathrm{tan}}.$$

The radial-only repair used in our experiments goes further. It retains only the radial curvature, which describes changes in distance, and applies damping in the tangential subspace:

$$\widetilde{\mathbf{H}}^{\mathrm{rad}}(\mathbf{v}) = [\lambda_{\mathrm{rad}}]_+\mathbf{P}_{\mathrm{rad}} + \epsilon_{\mathrm{tan}}\mathbf{P}_{\mathrm{tan}}.$$

Here $\epsilon_{\mathrm{tan}} \geq 0$ is a small tangential damping value, set to zero unless stated otherwise. This is a structured approximation, rather than the PSD projection of the Hessian, because it discards tangential curvature even when that curvature is positive.

For negative distance-Bernoulli pairs, write $\lambda_{\mathrm{dist,rad}} = 4(p/q^2)\|\mathbf{v}\|_2^2 - 2p/q$, where $p = \exp(-\|\mathbf{v}\|_2^2)$ and $q = 1 - p$. The radial-only repaired block is then

$$\widetilde{\mathbf{H}}_-^{\mathrm{dist,rad}}(\mathbf{v}) = [\lambda_{\mathrm{dist,rad}}]_+\mathbf{u}\mathbf{u}^\top + \epsilon_{\mathrm{tan}}\big(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top\big).$$

| Mode | Construction | What it does | What it does not do |
|---|---|---|---|
| Raw GGN | Use $\mathbf{H}_t = \nabla^2_{f_t} \ell_t$ directly | Retains the exact output Hessian of the linearized loss | Does not guarantee a PSD precision for distance or margin negatives |
| Spectral repair | Symmetrize, eigendecompose, and set negative eigenvalues to zero | Projects each output block onto the PSD cone, regardless of the loss | Removes negative curvature; does not turn it into positive evidence |
| Radial-only repair | Split distance-loss blocks into radial and tangential components; retain radial curvature and damp the tangential component | Gives a structured PSD approximation for distance losses; agrees with spectral repair on active negative margin blocks up to damping | Is not the PSD projection and can remove valid positive tangential curvature from already-PSD positive-pair blocks |
| Damping/jitter | Add isotropic precision after accumulation | Stabilizes inversion and can make the final matrix strictly positive definite | Does not identify or repair the loss-level curvature pathology |
| Empirical Fisher | Replace $\mathbf{H}_t$ by $\mathbf{g}_t \mathbf{g}_t^\top$ | Gives a PSD block by construction | Does not repair the Hessian and can reduce relational curvature to rank one per unit |

Table 3: Repair and fallback modes considered in this work. The main experiments use repaired GGN curvature, while empirical Fisher and alternative repairs are used as ablations.

For active margin negatives, the radial eigenvalue is 2 and the tangential eigenvalues are negative, giving

$$\widetilde{\mathbf{H}}_-^{\mathrm{mrg,rad}}(\mathbf{v}) = 2\mathbf{u}\mathbf{u}^\top + \epsilon_{\tan}\big(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top\big), \qquad 0 < \|\mathbf{v}\|_2 < m.$$

The loss is locally flat for inactive margin negatives away from the boundary. We set $\widetilde{\mathbf{H}}_-^{\mathrm{mrg,rad}}(\mathbf{v}) = \mathbf{0}$ for $\|\mathbf{v}\|_2 \geq m$, including the boundary convention described above.

The difference between spectral and radial-only repair is clearest for positive pairs. Both distance and margin losses have

$$\mathbf{H}_+^{\mathrm{dist}}(\mathbf{v}) = \mathbf{H}_+^{\mathrm{mrg}}(\mathbf{v}) = 2\mathbf{I}_d = 2\mathbf{P}_{\mathrm{rad}} + 2\mathbf{P}_{\tan}.$$

Spectral repair leaves these PSD blocks unchanged, so $\widetilde{\mathbf{H}}_+^{\mathrm{dist,spec}}(\mathbf{v}) = \widetilde{\mathbf{H}}_+^{\mathrm{mrg,spec}}(\mathbf{v}) = 2\mathbf{I}_d$. Radial-only repair instead retains the radial component and damps the tangential component:

$$\widetilde{\mathbf{H}}_+^{\mathrm{dist,rad}}(\mathbf{v}) = \widetilde{\mathbf{H}}_+^{\mathrm{mrg,rad}}(\mathbf{v}) = 2\mathbf{u}\mathbf{u}^\top + \epsilon_{\tan}\big(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top\big).$$

For active negative margin pairs, the two repairs therefore agree up to the small tangential damping term. For positive pairs, spectral repair preserves the full PSD Hessian $2\mathbf{I}_d$, while radial-only repair removes its valid positive tangential curvature. This makes radial-only repair more intrusive in Frobenius norm. We use it as an ablation or modeling choice that retains only evidence about changes in distance, rather than as the nearest PSD projection of the original Hessian. It applies specifically to radial distance losses and should not be used for candidate-logit losses such as InfoNCE.

**Damping and jitter are not curvature repairs.** Adding $\lambda\mathbf{I}$, or increasing the prior precision, can make the assembled parameter-space precision positive definite. In particular, for sufficiently large $\lambda$,

$$\mathbf{\Lambda}_0 + \sum_t \mathbf{J}_t^\top \mathbf{H}_t \mathbf{J}_t + \lambda\mathbf{I} \succ 0.$$

Although this can help numerically, it does not repair the contrastive output curvature: sufficiently large isotropic precision can simply overwhelm the negative-curvature blocks. In this work, we use damping and jitter to stabilize inversion after repairing the output curvature. They are not our primary means of obtaining a valid posterior from an indefinite loss geometry.

Logit-Bernoulli and InfoNCE give PSD output curvature directly, whereas negative pairs in distance-Bernoulli and pairwise margin losses can have negative tangential curvature. Spectral repair projects these indefinite blocks onto the PSD cone before they are accumulated in parameter space. Radial-only repair is a structured approximation for distance losses: it retains curvature associated with

changes in distance and damps the tangential, rotation-like directions. It can also remove positive tangential curvature, as it does for positive distance and margin pairs. Radial-only repair should therefore be understood as a modeling ablation that retains radial evidence, rather than as the nearest PSD projection of the Hessian.

## C Empirical Fisher Alternative for Relational Curvature

Our main construction uses the output Hessian $\mathbf{H}_t^{\mathrm{GGN}} = \nabla_{f_t}^2 \ell_t(f_t)$ in the relational GGN block $\mathbf{J}_t^\top \mathbf{H}_t^{\mathrm{GGN}} \mathbf{J}_t$. When the output Hessian is indefinite, as it is for negative distance-Bernoulli pairs or active negative margin pairs, we project it onto the PSD cone before accumulating the resulting blocks in parameter space. The empirical Fisher (EF) offers an alternative that does not require Hessian blocks.

For a relational unit $t$, let $\mathbf{g}_t = \nabla_{f_t} \ell_t(f_t)|_{f_t = f_t(\boldsymbol{\theta}^\star)}$ be the output gradient at the MAP solution. EF replaces the output Hessian with the outer product of this gradient:

$$\mathbf{H}_t^{\mathrm{EF}} = \mathbf{g}_t \mathbf{g}_t^\top. \tag{14}$$

The corresponding parameter-space precision is

$$\boldsymbol{\Sigma}_{\mathrm{EF}}^{-1} = \boldsymbol{\Lambda}_0 + \sum_t \mathbf{J}_t^\top \mathbf{H}_t^{\mathrm{EF}} \mathbf{J}_t = \boldsymbol{\Lambda}_0 + \sum_t (\mathbf{J}_t^\top \mathbf{g}_t)(\mathbf{J}_t^\top \mathbf{g}_t)^\top. \tag{15}$$

Each contribution is an outer product and is therefore PSD. With a strictly positive definite prior precision or positive damping, the resulting precision defines a proper Gaussian posterior approximation.

**Why EF is attractive.** EF has two practical advantages. It requires neither an eigendecomposition of the output Hessian nor a separate repair for distance or margin losses. It also applies in the same way to scalar scores, candidate-logit vectors, and embedding differences. This makes it a convenient fallback when the output Hessian is difficult to form, affected by nonsmoothness, or indefinite.

**EF is not a Hessian repair.** Although EF is PSD, it does not project the Hessian onto the PSD cone. Instead, it replaces second-order curvature with first-order sensitivity at the MAP solution, giving the posterior a different local interpretation. Spectrally repaired GGN retains the nonnegative curvature directions of the loss and removes the negative directions that cannot serve as precision. EF retains only the current gradient direction for each relational unit.

For relational contrastive objectives, this is a substantial difference. Each unit contributes a block of rank at most one, $\mathrm{rank}(\mathbf{H}_t^{\mathrm{EF}}) \leq 1$. The sum over many units can still have high rank, but an individual unit supplies only one direction. The GGN Hessian of a candidate-set loss such as InfoNCE, by comparison, can describe a full covariance structure over the candidates within that unit.

**Relation to true Fisher and GGN.** For well-specified likelihood models, and under the usual regularity conditions, the expected Fisher can coincide with the expected Hessian of the negative log-likelihood. The empirical Fisher in equation 14 instead uses the observed target for each training unit to form a single gradient outer product. For example, consider the scalar logit-Bernoulli loss $\ell(s, y) = \mathrm{softplus}(s) - ys$, with $p = \sigma(s)$. Its GGN curvature is $\mathbf{H}^{\mathrm{GGN}} = p(1-p)$, whereas its EF curvature is $\mathbf{H}^{\mathrm{EF}} = (p-y)^2$. Both are nonnegative, but they measure different quantities.

Similarly, for InfoNCE with candidate probabilities $\mathbf{q} = \mathrm{softmax}(f_t)$ and positive candidate $p(i)$, the two blocks are

$$\mathbf{H}_t^{\mathrm{GGN}} = \mathrm{diag}(\mathbf{q}) - \mathbf{q}\mathbf{q}^\top, \qquad \mathbf{H}_t^{\mathrm{EF}} = (\mathbf{q} - \mathbf{e}_{p(i)})(\mathbf{q} - \mathbf{e}_{p(i)})^\top.$$

The GGN block gives the softmax covariance over candidates. The EF block is an outer product of the current prediction error, with rank at most one, and therefore loses the multi-candidate interaction geometry retained by the GGN.

**Distance and margin losses.** The same distinction arises for distance-based losses. As before, write $\mathbf{v} = \mathbf{z}_i - \mathbf{z}_j$, $r = \|\mathbf{v}\|_2$, and $\mathbf{u} = \mathbf{v}/r$ for $r > 0$. A positive pair with squared-distance loss $\ell^+(\mathbf{v}) = \|\mathbf{v}\|_2^2$ has GGN/Hessian $\mathbf{H}_+^{\mathrm{GGN}} = 2\mathbf{I}_d$, but its EF block is $\mathbf{H}_+^{\mathrm{EF}} = (2\mathbf{v})(2\mathbf{v})^\top = 4\mathbf{v}\mathbf{v}^\top$.

Thus, even when the positive-pair Hessian is already PSD, EF replaces isotropic attraction curvature with a single radial direction.

For an active negative margin pair, the loss is $\ell_-^{\mathrm{mrg}}(\mathbf{v}) = (m - r)^2$, with $0 < r < m$. The raw Hessian has radial eigenvalue 2 and negative tangential eigenvalues. Both spectral and radial-only repair remove these negative tangential directions and retain the radial curvature, with radial-only repair optionally adding tangential damping:

$$\widetilde{\mathbf{H}}_{\mathrm{mrg}}^{\mathrm{GGN}} = 2\mathbf{u}\mathbf{u}^\top \quad \text{or} \quad 2\mathbf{u}\mathbf{u}^\top + \epsilon_{\tan}(\mathbf{I}_d - \mathbf{u}\mathbf{u}^\top).$$

Apart from this damping term, repaired GGN and EF are both PSD and radial, but they assign different precision along the radial direction. Repaired GGN retains the local second derivative of the active margin loss. EF uses the squared first derivative, so its precision vanishes as $r$ approaches the margin.

For a negative distance-Bernoulli pair, the loss is $\ell_-^{\mathrm{dist}}(\mathbf{v}) = -\log(1 - \exp(-\|\mathbf{v}\|_2^2))$. Writing $p = \exp(-r^2)$ and $q = 1 - p$, its gradient is $\nabla_{\mathbf{v}} \ell_-^{\mathrm{dist}} = -2(p/q)\mathbf{v}$, giving the EF block

$$\mathbf{H}_{\mathrm{dist}}^{\mathrm{EF}} = 4\frac{p^2}{q^2}\mathbf{v}\mathbf{v}^\top.$$

This block is again PSD and radial, but it is not the PSD projection of the Hessian. Spectral repair retains the positive radial Hessian eigenvalue and removes the negative tangential eigenvalues, whereas EF replaces the Hessian with a squared-gradient magnitude.

**What EF does and does not provide.** EF provides a simple PSD approximation, but it should not be taken to preserve the relational geometry of the contrastive loss. It produces PSD output blocks by construction, avoids explicit Hessian eigendecompositions and loss-specific repairs, and gives a simple fallback curvature for all relational units.

It does not, however, project the true Hessian onto the PSD cone or preserve high-rank candidate interactions such as the InfoNCE softmax covariance. It also discards tangential or isotropic curvature that may be present in PSD Hessian blocks, and it does not distinguish invalid negative curvature from valid positive curvature.

These properties make EF useful as a PSD fallback and as an ablation, but it changes the geometry of the local Bayesian model more substantially than spectral or radial repair. We therefore use repaired GGN curvature as the default in our main experiments.

We repeat the metagenomic binning experiment from Section 6.1, whose original results appear in Figure 2, using EF in place of GGN. Figure 3 shows the results.

## D Additional details on marginal variance and downstream uncertainty

In this section, we provide proofs for the propositions in Section 4. Also, to provide with visual intuition, we show how these results apply in a very simple example in Figure 4 and also for a pairwise score in Figure 5.

**Proposition 1.** Let $h_{\boldsymbol{\theta}}$ be a (linearized) contrastive encoder and $q(\boldsymbol{\theta})$ be a (repaired) LLA posterior, then, for any $\mathbf{x} \in \mathcal{X}$, $h_{\boldsymbol{\theta}}(\mathbf{x}) \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Consider a linear binary decision boundary defined as $D(\mathbf{z}) := \mathbf{1}\{\mathbf{a}^\top \mathbf{z} + b \geq 0\}$. Then, denoting the standard gaussian c.d.f. as $\Phi$, the posterior probability of changing the assignment made at the posterior mean is

$$\mathbb{P}[D(h(\mathbf{x})) \neq D(\boldsymbol{\mu}) \mid \mathcal{D}] = \Phi\left(-\frac{|\mathbf{a}^\top \boldsymbol{\mu} + b|}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}}}\right).$$

*Proof.* Since an affine transformation of a Gaussian vector is Gaussian,

$$\mathbf{a}^\top h_{\boldsymbol{\theta}}(\mathbf{x}) + b \mid \mathcal{D} \sim \mathcal{N}\left(\mathbf{a}^\top \boldsymbol{\mu} + b,\, \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}\right).$$

Write $m = \mathbf{a}^\top \boldsymbol{\mu} + b$ and $s^2 = \mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}$, and assume $s > 0$. If $m \geq 0$, then $D(\boldsymbol{\mu}) = 1$, so the assignment changes when the Gaussian score falls below zero. This has probability $\Phi(-m/s)$. If

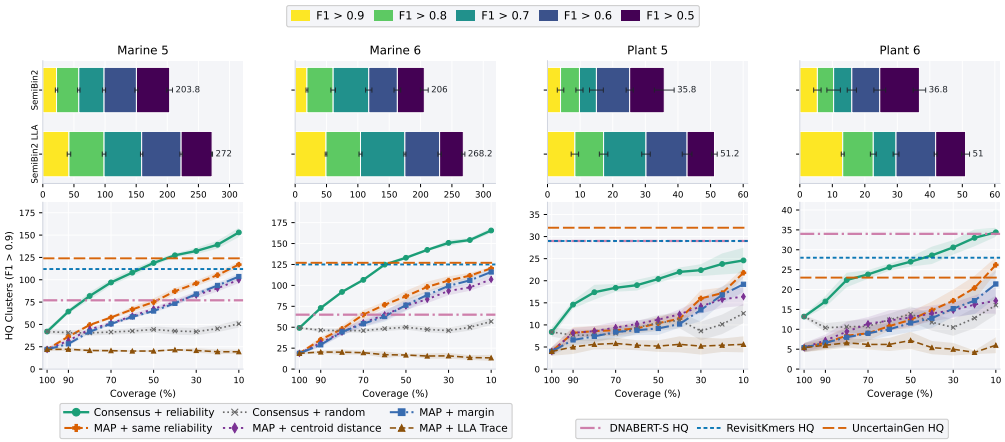



Figure 3: Genome reliable-core results on the Marine and Plant splits using EF in place of GGN, with $k = 500$ retained for k-Means. The top panels compare high-quality cluster counts across F1 thresholds for SemiBin2 and the LLA posterior-consensus variant. The bottom panels show the number of high-quality clusters retained as coverage decreases under the different approaches.

$m < 0$, then $D(\boldsymbol{\mu}) = 0$, and the assignment changes when the score is nonnegative, which has probability $1 - \Phi(-m/s) = \Phi(m/s)$.

In both cases, the disagreement probability is $\Phi(-|m|/s)$. Thus,

$$\mathbb{P}\left[D(h_{\boldsymbol{\theta}}(\mathbf{x})) \neq D(\boldsymbol{\mu}) \mid \mathcal{D}\right] = \Phi\left(-\frac{\left|\mathbf{a}^\top \boldsymbol{\mu} + b\right|}{\sqrt{\mathbf{a}^\top \boldsymbol{\Sigma} \mathbf{a}}}\right).$$

If $s = 0$, the score equals $m$ almost surely, so the disagreement probability is zero. □

**Proposition 2.** Suppose an LLA posterior $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma})$, the two posterior (linearized) scores are jointly Gaussian under

$$\begin{pmatrix} s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) \\ s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_k) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_{ij} \\ \mu_{ik} \end{pmatrix}, \begin{pmatrix} V_{ij} & C_{ij,ik} \\ C_{ik,ij} & V_{ik} \end{pmatrix}\right)$$

For the downstream comparison $A_i = \mathbf{1}\left\{s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) \geq s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_k)\right\}$. The probability of changing the score rank w.r.t the MAP is given by

$$\mathbb{P}\left[A_i(\boldsymbol{\theta}) \neq A_i(\boldsymbol{\theta}^\star)\right] = \Phi\left(-\frac{|\mu_{ij} - \mu_{ik}|}{\sqrt{V_{ij} + V_{ik} - 2C_{ij,ik}}}\right),$$

*Proof.* The two scores are jointly Gaussian, so their difference

$$\Delta_{\boldsymbol{\theta}} = s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_j) - s_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x}_k)$$

is Gaussian with mean $m = \mu_{ij} - \mu_{ik}$ and variance $v = V_{ij} + V_{ik} - 2C_{ij,ik}$.

The comparison is determined by the sign of this difference:

$$A_i(\boldsymbol{\theta}) = \mathbf{1}\{\Delta_{\boldsymbol{\theta}} \geq 0\}.$$

Since the linearized posterior is centered at $\boldsymbol{\theta}^\star$, the score difference at $\boldsymbol{\theta}^\star$ equals $m$, giving

$$A_i(\boldsymbol{\theta}^\star) = \mathbf{1}\{m \geq 0\}.$$

Assume $v > 0$. If $m \geq 0$, a rank change occurs when $\Delta_{\boldsymbol{\theta}} < 0$, with probability $\Phi(-m/\sqrt{v})$. If $m < 0$, it occurs when $\Delta_{\boldsymbol{\theta}} \geq 0$, with probability $1 - \Phi(-m/\sqrt{v}) = \Phi(m/\sqrt{v})$. Both cases therefore give

$$\mathbb{P}\left[A_i(\boldsymbol{\theta}) \neq A_i(\boldsymbol{\theta}^\star)\right] = \Phi\left(-\frac{|\mu_{ij} - \mu_{ik}|}{\sqrt{V_{ij} + V_{ik} - 2C_{ij,ik}}}\right).$$

If $v = 0$, the score difference equals $m$ almost surely, so the rank-change probability is zero. □

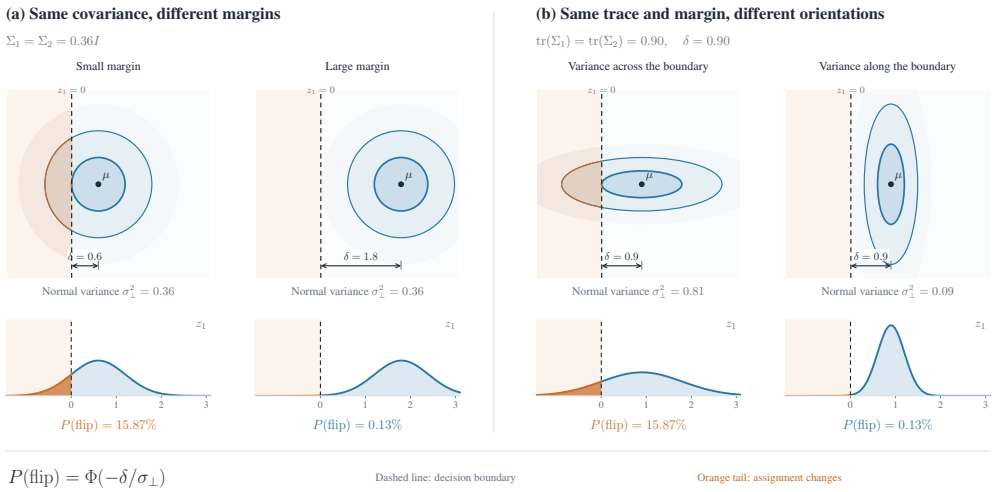



Figure 4: Marginal embedding variance does not determine assignment uncertainty. (a) With identical covariance matrices, a larger margin between the posterior mean and the decision boundary reduces the probability of changing assignment. (b) With identical means and covariance traces, variance perpendicular to the boundary produces greater assignment uncertainty than variance parallel to it. Upper panels show Gaussian embedding posteriors; lower panels show their projections onto the boundary-normal direction. Orange tails indicate assignment changes relative to the posterior mean.




Figure 5: Ranking uncertainty depends on the joint behavior of competing pairwise scores. (a) A query $x_i$ is compared with candidates $x_j$ and $x_k$, with both scores evaluated under the same posterior encoder draw. (b,c) The scores have identical marginal distributions, but positive covariance stabilizes their difference, whereas negative covariance increases ranking uncertainty. Each row connects the two scores from one illustrative draw; orange rows indicate reversals of the MAP ranking. Reported rank-flip probabilities are computed analytically from the joint Gaussian distribution.

# E EXPERIMENTAL AND REPRODUCIBILITY DETAILS

All experiments use the same posterior-propagation protocol. We first obtain a MAP contrastive encoder by loading a pretrained foundation/self-supervised encoder. We then fit a last-layer linearized Laplace approximation (LLA) around the MAP point, sample the last-layer parameters, and apply the downstream rule separately in each sampled embedding space. For clustering, the sampled predictions are combined through posterior coassignment probabilities; for classification, the sampled embeddings define either a Monte Carlo predictive readout or a Gaussian-logit readout. Reliable cores are selected only after this posterior propagation step.

Unless otherwise stated, experiments are repeated over five seeds, and reported aggregate values are means over seeds. The synthetic two-blob presented in the main text (see Figure 1) is an illustrative single-seed experiment. Ablations use the same architecture and posterior settings as the corresponding main experiment.

## E.1 DATASETS

**EMNIST Balanced and CIFAR-100.** We use EMNIST Balanced (Cohen et al., 2017), comprising 47 balanced character classes, and CIFAR-100 (Krizhevsky, 2009), comprising 100 object classes, for classification and curvature-repair ablations. EMNIST images are converted to the standard upright orientation and their grayscale channel is repeated three times. Images from both datasets are scaled to $[0, 1]$, resized to $224 \times 224$, and normalized using the standard CLIP channel-wise statistics. Unless otherwise stated, we use the complete training and test splits of both datasets.

**DNABERT-S metagenomic panels.** Genome experiments use the plant and marine panels from the DNABERT-S metagenomic benchmark (Zhou et al., 2024). We follow the SemiBin2 binning formulation (Pan et al., 2023): long contigs define must-link split pairs, and cannot-link pairs are sampled between contigs. The calibration split is the corresponding $\text{clustering}_0$ panel, while $\text{binning}_5$ and $\text{binning}_6$ are held-out transfer panels. The input representation is the 136-dimensional SemiBin2 k-mer feature vector; abundance and BAM coverage features are not used in this k-mer-only setting.

**SICAPv2 pathology.** SICAPv2 is a public prostate histopathology dataset derived from 155 biopsy whole-slide images from 95 patients (Silva-Rodríguez et al., 2020). The slides are provided as 18,783 RGB tissue patches of $512 \times 512$ pixels at $10\times$ magnification, extracted with 50% overlap; patches containing at most 20% tissue are excluded. Expert annotations identify non-cancerous tissue and Gleason patterns 3, 4, and 5 at the pixel level. For patch classification, cancerous patches are assigned the majority Gleason pattern in their annotation mask, while non-cancerous patches are drawn from slides classified as non-cancerous, yielding the four labels $\{\text{NC}, \text{G3}, \text{G4}, \text{G5}\}$. We use the official patient-separated test partition, comprising 9,959 training and 2,122 test patches, so patches from the same patient do not occur in both sets.

## E.2 ENCODERS AND LAST-LAYER TARGETS

Table 4 summarizes the architectures. In all cases, the LLA posterior is placed on a final linear map. For encoders taken from prior work, we use the published architecture and weights.

## E.3 TRAINING AND POSTERIOR FITTING

**Frozen-CLIP loss-and-repair ablation.** We study curvature repair on CIFAR-100 using a frozen pretrained CLIP ViT-B/32 visual encoder, without post-projection embedding normalization. KFAC GGN curvature is fitted on the final linear projection using either the distance-Bernoulli loss or the contrastive margin loss, each combined with spectral or radial-only repair. We accumulate at most 64 curvature batches of 16 supervised pairs, sampled with equal positive and negative proportions. Radial-only repair discards the tangential component, with $\epsilon_{\tan} = 0$. All four configurations are evaluated over five random seeds.

We reserve 50 images per class, excluding them from curvature fitting and classifier training in all four configurations. For the margin loss, $m$ is set to the 25th percentile of negative-pair distances on

Table 4: Frozen encoder architectures and linear layers targeted by LLA. No feature extractor is trained or fine-tuned in any reported experiment.

| Setting | Frozen encoder | LLA target | Embedding used downstream |
|---|---|---|---|
| CLIP image experiments and ablations | Pretrained CLIP ViT-B/32 visual tower (Radford et al., 2021), including its learned visual projection. Post-projection normalization is not applied before the LLA target. | CLIP visual projection, a linear map from hidden width 768 to projection dimension 512. | 512-dimensional CLIP visual-projection embedding. |
| SimCLR image experiments | Pretrained ImageNet SimCLRv2 ResNet-50 $1\times$, SK0 checkpoint (Chen et al., 2020b). Inputs are resized to $224 \times 224$, and the converted ResNet projection output is retained as the LLA target. | Final ResNet $fc$ layer. | Output of the final linear layer. |
| SemiBin2 genome experiments | Pretrained SemiBin2 global-environment encoder (Pan et al., 2023), evaluated on 136-dimensional k-mer features. | Last linear layer of the SemiBin2 encoder. | SemiBin2 contig embedding. |
| SICAPv2/CONCH pathology | Pretrained CONCH-v1.5 vision encoder (Lu et al., 2024). The LLA implementation exposes the final attention-pooling output map before the contrastive layer normalization. | Final linear output map of the CONCH attention pool. | CONCH patch embedding before the final normalization used by the adapter. |

this held-out subset. The same validation pairs and selected margin are used for both repair modes; the margin is not selected by classification performance.

For posterior-predictive classification, the prior precision is selected from $\{1, 10^2, 10^4\}$ jointly with the logistic-classifier regularization parameter $C \in \{1, 10^3, 10^4\}$. Selection uses a second, disjoint balanced validation subset of 50 images per class drawn from the remaining training images. For each value of $C$, a logistic classifier is fitted to the MAP embeddings of the corresponding training partition and held fixed while each candidate prior is evaluated using four posterior samples. Candidate pairs are ranked by validation negative log-likelihood. After selection, we draw 64 posterior samples. For each sample, we compute the corresponding training and test embeddings, fit a separate logistic classifier using the selected $C$, and obtain test-set probabilities. Final posterior-predictive probabilities are computed by averaging across the 64 independently fitted classifiers.

**CLIP and SimCLR image experiments.** The pretrained image backbones are kept frozen, and KFAC factors are fitted on supervised InfoNCE candidate-set batches from the training split. The candidate set size is 10, the curvature batch size is 64, and at most 64 curvature batches are accumulated. We use temperature $0.07$ for CLIP and $0.1$ for SimCLR, and draw 16 posterior samples for the final evaluation. For classification, the prior precision is selected from $\{1, 10^2, 10^4\}$, jointly with $C \in \{1, 10^3, 10^4\}$, using a balanced training-only validation subset. Each candidate is evaluated using four posterior samples and selected primarily by validation negative log-likelihood. For clustering, the prior precision is selected from $\{10^3, 10^4, 10^5, 10^6\}$ using the number of high-quality clusters obtained by posterior-consensus clustering on the training split, with macro-F1 and Hungarian-aligned accuracy as tie-breakers. We evaluate both empirical-Fisher KFAC and GGN KFAC. Neither configuration requires a curvature repair because the InfoNCE output Hessian is positive semidefinite.

**SemiBin2 genome experiments.** The pretrained SemiBin2 encoder is kept fixed, and KFAC is fitted on its final encoder linear layer. Each KFAC fit uses 64 curvature batches, a positive batch size of 512, one randomly sampled cannot-link pair per positive pair, and margin $m = 1$. The prior-precision grid is $\{10^3, 10^4, 10^5, 10^6\}$. Prior selection is performed separately for each source dataset, random seed, and KFAC configuration on the corresponding `clustering_0` calibration panel. The selected posterior is then transferred unchanged to `binning_5` and `binning_6`. The experiments

use 20 K-means initializations, and draw 32 posterior samples. Because the SemiBin2 objective is a vector-difference margin loss, the GGN configurations use either spectral repair or radial-only repair; empirical-Fisher KFAC with no repair is included as a positive semidefinite alternative. Also, following Metabat and DNABERT (Zhou et al., 2024; Kang et al., 2015), we only keep sequences longer than 2500 bp and we filter out species with fewer than 10 sequences.

**SICAPv2 pathology classification.** The CLIP and CONCH-v1.5 pathology backbones are kept frozen, and KFAC is fitted on the corresponding final linear projection. CLIP is evaluated at $224 \times 224$ resolution using CLIP normalization, whereas CONCH-v1.5 is evaluated at $448 \times 448$ resolution using ImageNet normalization. Both models use supervised InfoNCE candidate sets of size 8 and accumulate at most 256 curvature batches. The curvature batch size is 16 for CLIP and 4 for CONCH; the corresponding temperatures are $0.07$ and $0.1$. The prior precision is selected from $\{1, 10^2, 10^4\}$, jointly with $C \in \{1, 10^3, 10^4\}$, on a balanced training-only validation subset using four posterior samples per candidate. Final classification results use 16 posterior samples. We evaluate empirical-Fisher KFAC and GGN KFAC without repair, with damping $10^{-4}$ and jitter $10^{-6}$.

### E.4 Downstream Rules and Metrics

**Clustering.** MAP clustering applies K-means in the MAP embedding space with $K = 500$ except for the K-ablation, where we vary this value. Posterior clustering draws embeddings from the LLA posterior, reruns the same K-means rule for each draw, and forms a posterior-consensus clustering from the coassignment matrix.

Cluster labels are evaluated by Hungarian alignment to ground-truth labels. We report mainly the number of High-Quality clusters in the experimental section, and also macro-F1 and ARI in the additional Genome Curvature comparison (see Figure 8). A high-quality cluster is one whose aligned retained-set F1 exceeds the stated threshold, usually 0.9. Figure-style stacked bars count clusters with F1 greater than 0.5, 0.6, 0.7, 0.8, and 0.9.

**Classification.** For image classification, the MAP readout is multinomial logistic regression trained on MAP embeddings. The sampled-LLA readout trains the same logistic classifier in each sampled embedding space and averages predictive probabilities. The Gaussian-logit readout trains a single downstream logistic model against the Gaussian embedding distribution induced by the last-layer posterior, then averages predictive probabilities under that distribution. Classification metrics are top-1 accuracy, negative log likelihood, and 15-bin top-label expected calibration error.

**Reliable cores.** For clustering, posterior reliability for an item is the average posterior agreement between that item and the other items it has been clustered with. Reliable-core curves retain either the globally highest-reliability items or the highest-reliability items within each predicted cluster. The main coverage curves use the per-cluster rule. Baselines include pointwise embedding-trace rejection, MAP centroid similarity for genome coverage curves, and repeated random rejection.

## F Additional Clustering Experimental Results

We provide additional metagenomic binning clustering results performed on the main text of the paper. Tables 5 and 6 provide numerical results to Figure 2. Figures 6 and 7 show the results of the task when using $K = 100$ and $K = 200$, respectively.

Figure 8 isolates the differences between the curvature configurations and repairs. The figure shows the difference between the results obtained by GGN with spectral repair, with EF and GGN with radial repair, respectively. Notice that most mean differences lie close to zero: absolute Macro F1 differences remain below 0.33 percentage points, ARI differences remain below 0.69 points on the 0–100 scale, and full-coverage HQ differences are at most three clusters. Mean maximum HQ counts differ by at most two clusters. These results indicate broadly comparable clustering outcomes across the configurations, with the largest differences occurring in full-coverage HQ counts on Marine 6.

Table 5: Numbers of clusters exceeding each aligned F1 threshold for direct SemiBin2 MAP assignments and posterior-consensus solutions, using $K = 500$ from the cluster-count ablation. Values are mean ± standard deviation over five seeds.

| Method | F1 > 0.9 | F1 > 0.8 | F1 > 0.7 | F1 > 0.6 | F1 > 0.5 |
|---|---|---|---|---|---|
| **Marine 5** | | | | | |
| MAP | $20.4 \pm 1.1$ | $56.8 \pm 2.2$ | $97.6 \pm 3.4$ | $150.8 \pm 3.4$ | $204.8 \pm 5.3$ |
| Consensus, EF | $42.0 \pm 2.8$ | $98.4 \pm 2.9$ | $158.8 \pm 1.9$ | $222.2 \pm 1.9$ | $272.0 \pm 0.7$ |
| Consensus, GGN/radial | $42.8 \pm 1.5$ | $97.0 \pm 1.4$ | $156.4 \pm 1.9$ | $217.2 \pm 2.7$ | $268.2 \pm 3.2$ |
| Consensus, GGN/spectral | $42.6 \pm 2.1$ | $97.8 \pm 1.9$ | $159.0 \pm 0.7$ | $220.8 \pm 3.7$ | $268.2 \pm 2.3$ |
| **Marine 6** | | | | | |
| MAP | $18.6 \pm 0.9$ | $59.4 \pm 1.5$ | $114.2 \pm 5.1$ | $161.2 \pm 6.0$ | $204.0 \pm 7.0$ |
| Consensus, EF | $49.2 \pm 1.3$ | $104.2 \pm 2.4$ | $175.0 \pm 1.6$ | $230.4 \pm 2.7$ | $268.2 \pm 2.5$ |
| Consensus, GGN/radial | $50.4 \pm 3.5$ | $106.0 \pm 4.8$ | $176.6 \pm 3.0$ | $229.8 \pm 2.3$ | $269.8 \pm 3.0$ |
| Consensus, GGN/spectral | $47.4 \pm 1.7$ | $104.2 \pm 2.9$ | $174.4 \pm 3.5$ | $230.0 \pm 2.5$ | $267.8 \pm 1.9$ |
| **Plant 5** | | | | | |
| MAP | $4.4 \pm 1.1$ | $10.2 \pm 1.5$ | $16.0 \pm 2.0$ | $24.2 \pm 0.8$ | $36.6 \pm 2.7$ |
| Consensus, EF | $8.4 \pm 1.1$ | $17.2 \pm 1.3$ | $30.2 \pm 1.3$ | $42.8 \pm 1.5$ | $51.2 \pm 0.8$ |
| Consensus, GGN/radial | $9.2 \pm 0.8$ | $17.6 \pm 1.1$ | $28.8 \pm 1.1$ | $42.6 \pm 2.6$ | $51.8 \pm 2.5$ |
| Consensus, GGN/spectral | $8.8 \pm 1.3$ | $18.6 \pm 0.5$ | $31.2 \pm 1.3$ | $42.0 \pm 1.2$ | $52.0 \pm 1.6$ |
| **Plant 6** | | | | | |
| MAP | $4.6 \pm 1.9$ | $10.0 \pm 2.3$ | $15.4 \pm 0.9$ | $24.2 \pm 1.3$ | $36.2 \pm 2.6$ |
| Consensus, EF | $13.2 \pm 0.4$ | $22.6 \pm 0.9$ | $29.6 \pm 0.9$ | $41.8 \pm 0.4$ | $51.0 \pm 1.2$ |
| Consensus, GGN/radial | $13.2 \pm 0.4$ | $23.4 \pm 0.5$ | $30.0 \pm 1.2$ | $41.4 \pm 1.1$ | $50.8 \pm 2.0$ |
| Consensus, GGN/spectral | $13.0 \pm 0.7$ | $23.0 \pm 1.4$ | $29.6 \pm 1.1$ | $40.8 \pm 1.3$ | $50.2 \pm 0.8$ |

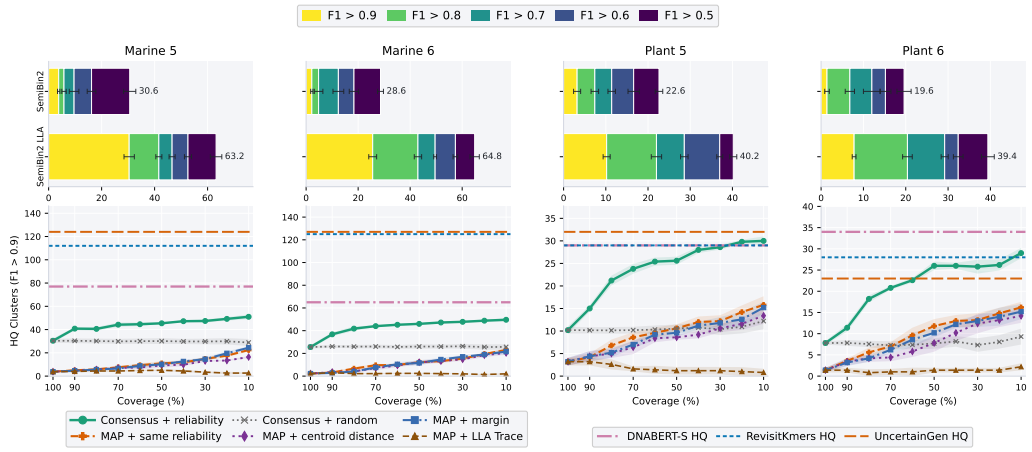



Figure 6: Genome reliable-core results on Marine and Plant splits, using $K = 100$ in k-Means, and using the GGN approximation with spectral repair. Top panels compare high-quality cluster counts for SemiBin2 and the LLA posterior-consensus variant across F1 thresholds. Bottom panels show high-quality clusters retained as coverage is reduced with different approaches.

## G Ablation Studies

**Number of posterior samples.** The sample-count ablation varies

$$S \in \{1, 2, 4, 8, 16, 32, 64\}.$$

For images, this grid is run over both datasets, both pretrained backbones, see the results in Figure 9.

Increasing the number of posterior samples primarily improves predictive uncertainty rather than classification accuracy. On CIFAR-100, NLL and ECE improve consistently as more samples are

Table 6: Numbers of high-quality clusters (aligned F1 $> 0.9$) retained at selected coverage levels under consensus and MAP selection rules, using $K = 500$ from the cluster-count ablation. Values are mean $\pm$ standard deviation over five seeds.

| Selector | 90% | 70% | 50% | 30% | 10% |
|---|---|---|---|---|---|
| **Marine 5** | | | | | |
| Consensus + rel. (EF) | $64.4 \pm 2.3$ | $97.0 \pm 3.4$ | $118.8 \pm 3.1$ | $132.2 \pm 3.6$ | $153.0 \pm 5.2$ |
| Consensus + rel. (GGN/radial) | $63.6 \pm 2.3$ | $96.0 \pm 3.1$ | $114.6 \pm 2.3$ | $133.2 \pm 4.4$ | $152.6 \pm 8.6$ |
| Consensus + rel. (GGN/spectral) | $65.0 \pm 1.4$ | $95.6 \pm 3.1$ | $118.4 \pm 1.9$ | $136.0 \pm 2.3$ | $154.0 \pm 3.7$ |
| Consensus + random (EF) | $40.8 \pm 2.8$ | $41.5 \pm 2.4$ | $44.3 \pm 3.1$ | $42.0 \pm 2.9$ | $50.7 \pm 2.0$ |
| Consensus + random (GGN/radial) | $41.1 \pm 0.8$ | $42.1 \pm 1.3$ | $44.3 \pm 1.4$ | $42.5 \pm 0.9$ | $50.8 \pm 1.5$ |
| Consensus + random (GGN/spectral) | $40.5 \pm 1.4$ | $41.2 \pm 1.7$ | $44.6 \pm 1.3$ | $41.9 \pm 1.6$ | $50.4 \pm 1.4$ |
| MAP + same reliability | $34.6 \pm 3.8$ | $56.6 \pm 1.5$ | $76.2 \pm 2.2$ | $99.8 \pm 4.9$ | $120.2 \pm 3.1$ |
| MAP + centroid distance | $30.2 \pm 2.3$ | $49.2 \pm 3.3$ | $66.8 \pm 2.4$ | $84.0 \pm 3.4$ | $99.2 \pm 1.1$ |
| MAP + margin | $27.6 \pm 2.1$ | $50.8 \pm 2.2$ | $65.0 \pm 2.5$ | $84.6 \pm 2.4$ | $105.4 \pm 8.4$ |
| MAP + LLA Trace | $20.8 \pm 1.1$ | $20.0 \pm 1.9$ | $19.6 \pm 1.5$ | $18.2 \pm 1.6$ | $16.4 \pm 1.1$ |
| **Marine 6** | | | | | |
| Consensus + rel. (EF) | $73.0 \pm 1.9$ | $106.6 \pm 1.3$ | $133.0 \pm 2.0$ | $150.8 \pm 3.6$ | $165.6 \pm 3.3$ |
| Consensus + rel. (GGN/radial) | $76.2 \pm 1.6$ | $110.8 \pm 3.3$ | $135.6 \pm 3.0$ | $153.2 \pm 3.2$ | $165.8 \pm 5.7$ |
| Consensus + rel. (GGN/spectral) | $75.4 \pm 2.3$ | $108.4 \pm 4.0$ | $132.2 \pm 2.6$ | $150.2 \pm 2.2$ | $165.8 \pm 5.8$ |
| Consensus + random (EF) | $46.5 \pm 1.9$ | $46.0 \pm 1.6$ | $50.0 \pm 1.3$ | $46.0 \pm 1.3$ | $56.9 \pm 1.7$ |
| Consensus + random (GGN/radial) | $47.5 \pm 2.6$ | $46.9 \pm 2.2$ | $51.0 \pm 2.7$ | $47.6 \pm 1.6$ | $57.6 \pm 2.0$ |
| Consensus + random (GGN/spectral) | $44.7 \pm 1.8$ | $44.5 \pm 1.9$ | $48.6 \pm 2.4$ | $45.0 \pm 1.8$ | $55.1 \pm 2.1$ |
| MAP + same reliability | $34.4 \pm 2.6$ | $62.8 \pm 2.4$ | $89.2 \pm 5.4$ | $103.4 \pm 3.4$ | $119.8 \pm 2.7$ |
| MAP + centroid distance | $32.0 \pm 2.7$ | $54.2 \pm 3.5$ | $72.0 \pm 3.2$ | $92.0 \pm 3.7$ | $106.8 \pm 3.3$ |
| MAP + margin | $30.2 \pm 3.0$ | $53.2 \pm 1.3$ | $76.0 \pm 4.4$ | $96.2 \pm 2.7$ | $112.6 \pm 3.2$ |
| MAP + LLA Trace | $19.8 \pm 1.1$ | $18.2 \pm 0.8$ | $16.8 \pm 1.1$ | $15.8 \pm 1.5$ | $14.4 \pm 3.8$ |
| **Plant 5** | | | | | |
| Consensus + rel. (EF) | $14.6 \pm 0.9$ | $18.4 \pm 0.9$ | $20.4 \pm 0.9$ | $22.4 \pm 1.3$ | $24.6 \pm 2.9$ |
| Consensus + rel. (GGN/radial) | $14.2 \pm 1.3$ | $18.2 \pm 0.8$ | $21.0 \pm 1.7$ | $22.6 \pm 0.9$ | $24.2 \pm 1.3$ |
| Consensus + rel. (GGN/spectral) | $15.4 \pm 1.7$ | $19.0 \pm 1.0$ | $21.6 \pm 1.5$ | $22.8 \pm 1.8$ | $26.2 \pm 1.3$ |
| Consensus + random (EF) | $7.7 \pm 0.7$ | $8.1 \pm 1.1$ | $10.4 \pm 1.0$ | $8.6 \pm 1.2$ | $12.6 \pm 1.0$ |
| Consensus + random (GGN/radial) | $7.9 \pm 0.8$ | $8.3 \pm 0.8$ | $11.0 \pm 0.9$ | $7.8 \pm 0.5$ | $13.1 \pm 0.7$ |
| Consensus + random (GGN/spectral) | $8.5 \pm 1.1$ | $8.8 \pm 1.0$ | $11.5 \pm 1.0$ | $8.4 \pm 0.8$ | $13.0 \pm 0.8$ |
| MAP + same reliability | $8.8 \pm 1.8$ | $9.4 \pm 1.5$ | $11.4 \pm 1.7$ | $15.8 \pm 1.6$ | $21.8 \pm 2.8$ |
| MAP + centroid distance | $9.0 \pm 1.6$ | $9.8 \pm 2.2$ | $11.6 \pm 0.9$ | $14.4 \pm 0.9$ | $16.4 \pm 1.1$ |
| MAP + margin | $6.0 \pm 1.2$ | $8.8 \pm 1.9$ | $10.0 \pm 1.4$ | $13.6 \pm 2.5$ | $19.6 \pm 2.4$ |
| MAP + LLA Trace | $4.8 \pm 1.5$ | $6.0 \pm 1.4$ | $5.6 \pm 1.5$ | $5.8 \pm 1.5$ | $5.8 \pm 1.8$ |
| **Plant 6** | | | | | |
| Consensus + rel. (EF) | $17.0 \pm 1.4$ | $23.8 \pm 1.1$ | $27.0 \pm 1.2$ | $30.6 \pm 1.8$ | $34.4 \pm 1.1$ |
| Consensus + rel. (GGN/radial) | $17.2 \pm 0.8$ | $25.6 \pm 0.9$ | $27.6 \pm 0.9$ | $32.0 \pm 1.6$ | $35.0 \pm 1.9$ |
| Consensus + rel. (GGN/spectral) | $17.4 \pm 1.5$ | $25.6 \pm 0.5$ | $27.6 \pm 0.9$ | $31.8 \pm 1.8$ | $35.4 \pm 1.1$ |
| Consensus + random (EF) | $10.4 \pm 0.7$ | $10.8 \pm 0.9$ | $13.8 \pm 0.9$ | $10.5 \pm 0.5$ | $16.1 \pm 0.4$ |
| Consensus + random (GGN/radial) | $10.6 \pm 1.1$ | $11.2 \pm 1.0$ | $14.3 \pm 0.3$ | $11.0 \pm 0.4$ | $16.2 \pm 0.3$ |
| Consensus + random (GGN/spectral) | $10.1 \pm 1.4$ | $10.5 \pm 1.1$ | $14.1 \pm 0.6$ | $11.1 \pm 0.5$ | $16.2 \pm 0.5$ |
| MAP + same reliability | $6.0 \pm 1.9$ | $8.4 \pm 2.8$ | $12.6 \pm 2.1$ | $18.2 \pm 1.6$ | $25.8 \pm 2.3$ |
| MAP + centroid distance | $6.8 \pm 1.8$ | $11.0 \pm 2.0$ | $12.8 \pm 1.9$ | $15.8 \pm 2.8$ | $18.2 \pm 3.6$ |
| MAP + margin | $6.0 \pm 1.9$ | $8.2 \pm 2.2$ | $11.8 \pm 2.3$ | $17.0 \pm 2.0$ | $22.4 \pm 2.3$ |
| MAP + LLA Trace | $5.2 \pm 2.2$ | $5.6 \pm 2.1$ | $6.4 \pm 2.5$ | $4.8 \pm 1.6$ | $5.2 \pm 1.8$ |

averaged, with most of the gain obtained by approximately 16 or 32 samples and only modest changes thereafter. Accuracy also improves slightly once multiple samples are used. On EMNIST, the MAP classifier is already strong and posterior averaging produces comparatively small changes, but NLL and ECE generally remain improved. The empirical-Fisher and GGN results exhibit similar convergence patterns.

For genome binning, the same grid is run for the SemiBin2 Plant/Marine fit-transfer protocol. In this case, since the margin loss is used, we present the results using spectral and radial repairs, and also empirical fisher. The results of this experiments are presented in Figure 10. In this experiment, we also observe an improvement in the metrics when we increase the number of posterior samples, stabilizing in most of the cases when using 8 samples.

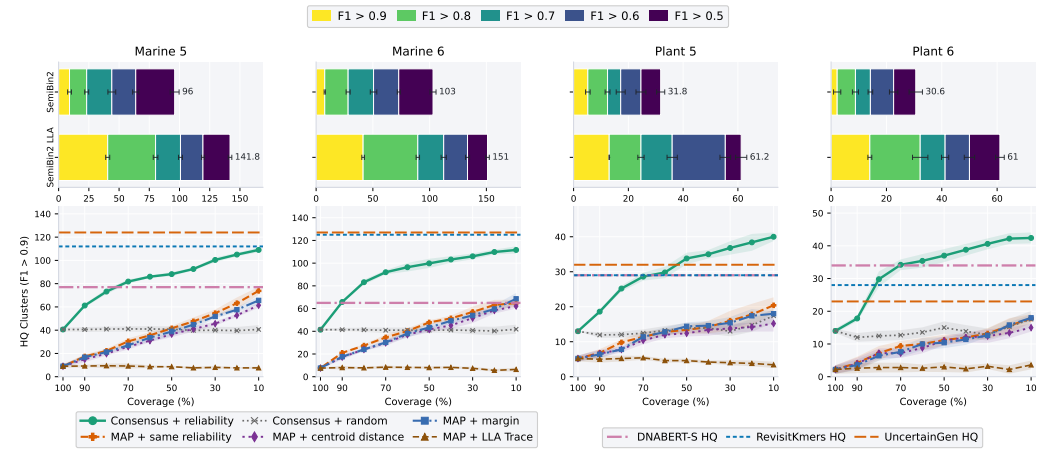



Figure 7: Genome reliable-core results on Marine and Plant splits, using $K = 200$ in k-Means, and using the GGN approximation with spectral repair. Top panels compare high-quality cluster counts for SemiBin2 and the LLA posterior-consensus variant across F1 thresholds. Bottom panels show high-quality clusters retained as coverage is reduced with different approaches.

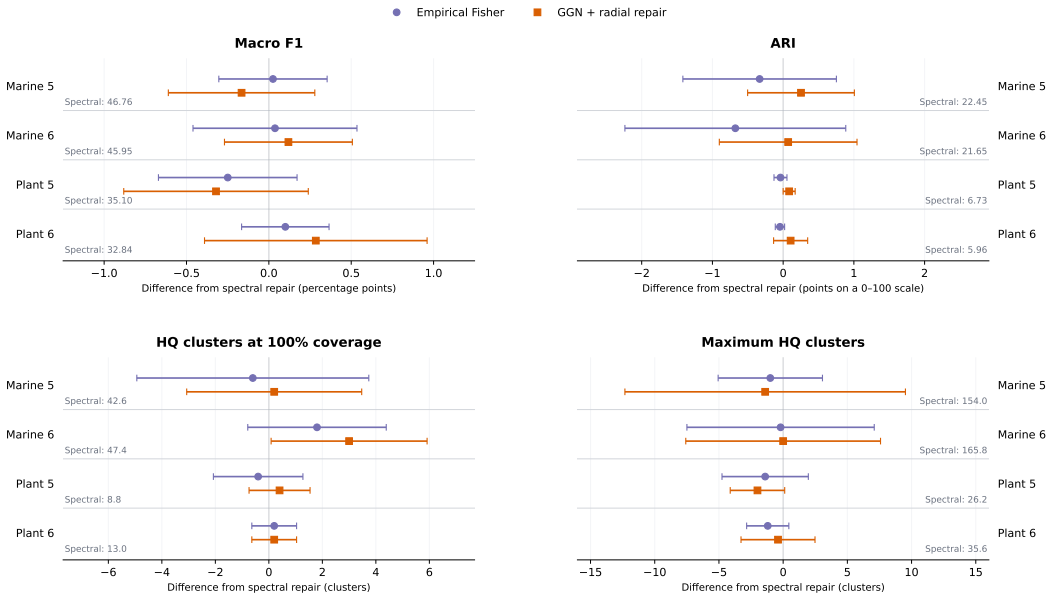



Figure 8: Genome curvature comparison at $K = 500$, using GGN with spectral repair as the reference. Points show mean paired differences (alternative minus spectral) across five seeds; error bars indicate one standard deviation of those differences. Macro F1 and ARI differences are expressed on a 0–100 scale, and HQ differences are expressed in cluster counts. Maximum HQ counts are computed separately for each seed over the 10–100% coverage grid using consensus-based reliability selection before taking paired differences.

**Number of clusters in genome K-means.** The K-ablation is restricted to SemiBin2 genome K-means. Instead of using oracle $K$, we evaluate

$$K \in \{100, 200, 500\}.$$

The posterior sample count remains 32. This ablation tests whether posterior reliability and high-quality-cluster counts remain stable when the downstream clustering rule is moved away from the true number of reference labels. The results, displayed in Figure 11, show that while in the Marine

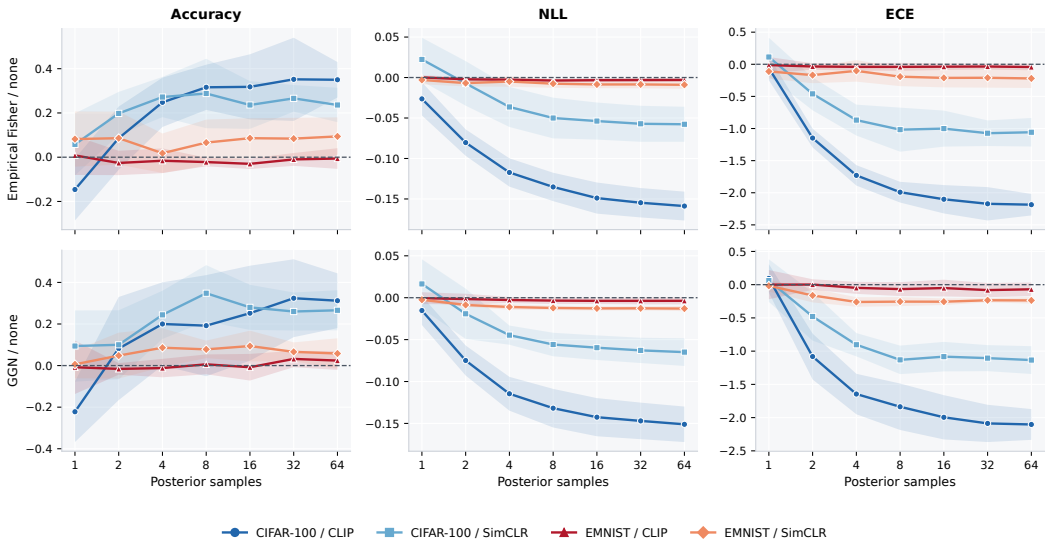



Figure 9: Image classification posterior-sample ablation. We report the results as the difference between the performance of our proposal and that of the MAP solution. No repair step is required in these experiments, as the InfoNCE loss is used.

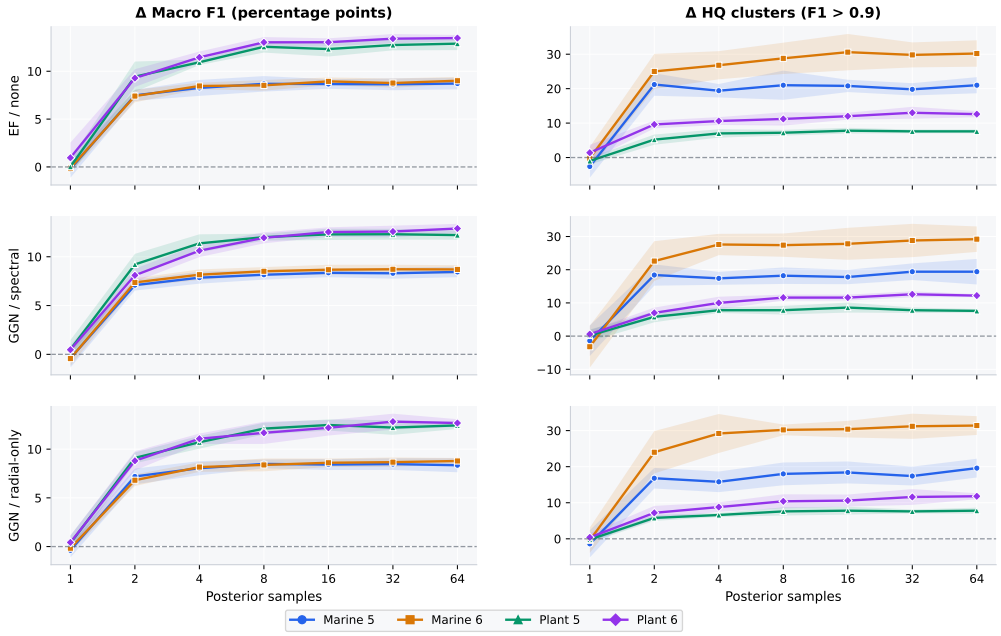



Figure 10: SemiBin2 posterior-sample ablation. We report the results as the difference between the performance of our proposal and that of the MAP solution.

splits increasing the number of clusters $K$ leads to an improvement in the number of HQ clusters, they do not improve in the Plant splits when increasing $K$ from 200 to 500.

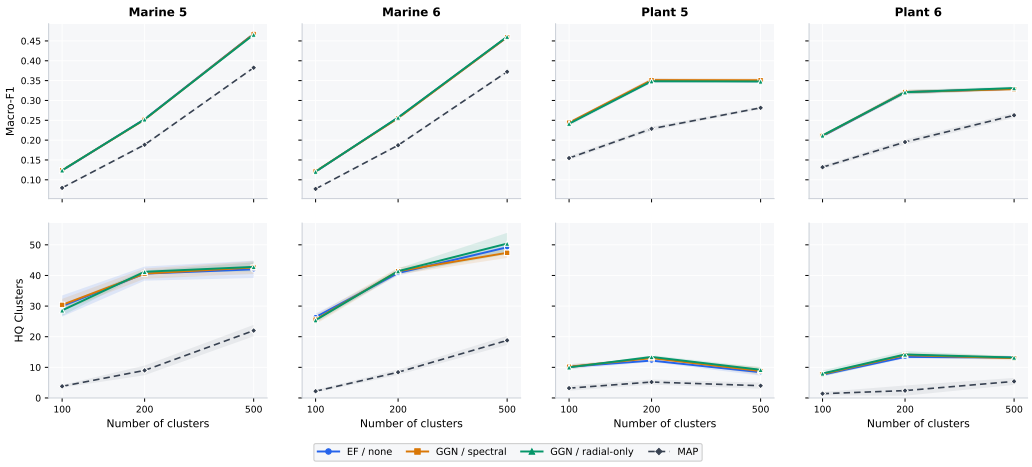



Figure 11: SemiBin2 clustering-$K$ ablation. Values are mean $\pm$ one standard deviation over five seeds at full coverage. MAP denotes standard SemiBin2 clustering and is shown once per $K$ using the EF/none run outputs.