CREDAL WRAPPER FOR UNCERTAINTY MODELLING IN SPARSE GEOLOGICAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models for mineral exploration are trained on a few hundred sparse, expensive samples and used to decide where to drill next, so their epistemic uncertainty (EU), the part that more data could remove, matters as much as their predictions. Yet EU cannot be measured on real data, because the counterfactual that defines it is never observed, so estimators are graded by proxies such as out-of-distribution detection and calibration. We contribute a general simulation protocol that makes EU directly computable for a given model and can grade any EU estimator. It separates two estimands the literature conflates, reducible error and estimator spread, builds an oracle reference for each in a synthetic prospectivity field sampled by clustered drilling surveys, and compares uncertainty measures of an identical ensemble within each simulated world. We use it to ask when credal (imprecise-probability) measures are more faithful than Bayesian model averaging (BMA) in the sparse, low-class regime of geological classification. Three findings follow. First, the credal interval width recovers the epistemic ordering better than BMA when members are few ($\Delta\rho \approx +0.06$ to $+0.07$ in rank correlation for $M \leq 5$, $P < 10^{-4}$), confirming the small-ensemble advantage reported for credal wrappers; the advantage vanishes by $M \approx 10$ and reverses at $M = 40$. Second, the choice of credal functional matters more than the choice between credal and BMA: in binary (deposit versus barren) classification the entropy interval carries no epistemic signal ($\rho \approx 0.00$) while the interval width of the same credal set does. Third, the credal advantage does not widen as training data are thinned, so it compensates for few posterior samples rather than few data. On real drill-hole geochemistry from South Australia, a resampling oracle reproduces the first finding. The protocol applies to any EU estimator whose data-generating process can be simulated.

1 INTRODUCTION

Epistemic uncertainty (EU), the part of predictive uncertainty that more data could remove, drives active learning, selective prediction and sequential experimental design (Kendall & Gal, 2017; Hüllermeier & Waegeman, 2021; Settles, 2009). It matters most where data are scarce and decisions are expensive. Mineral exploration is a clear example: models trained on a few hundred irregularly placed drill samples are used to choose where to drill next; the new results update the model, which then chooses the following hole (Mern et al., 2024; Rodriguez-Galiano et al., 2018). The value of this sequential cycle depends on whether the model's EU points to locations where new data would actually teach it something. In such settings ensembles are also small, because every ensemble member or posterior sample is an expensive fit.

Most practical EU estimators read uncertainty from a finite set of member predictions, whether deep-ensemble members, posterior samples of a Bayesian neural network, or dropout masks (Lakshminarayanan et al., 2017; Gal & Ghahramani, 2016; Jospin et al., 2022); we call any such set an *ensemble*. The standard approach is Bayesian model averaging (BMA), with epistemic uncertainty measured by the mutual-information decomposition (Houlsby et al., 2011). Credal wrappers (Wang et al., 2025a) instead retain the range of member predictions as a credal set, a convex set of distributions grounded in imprecise probability (Cuzzolin, 2020), and credal-set width has been used directly as an EU measure (Manchingal, 2025). Credal wrappers were shown to give more reliable uncertainty than BMA when only a few samples are available. We use the post-hoc credal wrapper

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.

because it applies to any existing ensemble without retraining; our conclusions apply to any method that outputs per-class probability intervals.

Any claim of this kind faces a basic measurement problem. EU is defined by a counterfactual: how would the prediction change if the model had been trained on different data of the same kind? On real data that counterfactual is never observed, so EU has no reference value and is evaluated through proxies such as out-of-distribution detection and calibration. These proxies are informative but do not isolate the epistemic estimate, and recent work argues that evaluation should target a quantity such as the reducible error of the deployed predictor (Paplhám et al., 2026; Jiménez et al., 2026). A second, quieter problem is that "epistemic uncertainty" names at least two different quantities: the excess error that more data would remove (Hüllermeier & Waegeman, 2021; Lahlou et al., 2023), and the spread of the estimator under resampling (Houlsby et al., 2011; Depeweg et al., 2018). They usually agree and diverge exactly when all members are confidently wrong together.

For example, if a survey never samples a barren granite next to a drilled ore body, every ensemble member extrapolates the mineralisation into the granite in the same way: the members agree, so the spread is near zero, yet all are wrong, so the reducible error is large. An ensemble trained on scans from one hospital and applied to another scanner fails in the same way. Disagreement-based estimators, BMA and credal alike, cannot see such cases.

We address both problems with an evaluation protocol in which the data-generating process is known, so each EU estimand has a computable oracle reference for a given model (Figure 1). The protocol is general: it can grade any estimator that assigns an EU score to each input, whether BMA, MC-Dropout, an evidential model or a credal set, provided the data-generating process can be simulated. We apply it to a question on which the evidence so far has come only from proxies: under which conditions, and through which functional, does a credal uncertainty measure of a finite ensemble recover EU more faithfully than BMA? The paper is a methods paper with a geoscience case study. Geology sets the regime we study (few ensemble members, few classes, small and spatially clustered datasets) and supplies the real-data test.

Novelty. Credal and BMA uncertainty measures have so far been compared through proxies such as out-of-distribution detection, calibration and accuracy (Wang et al., 2025a; Shaker & Hüllermeier, 2021). Oracles from a known data-generating process have measured the reducible error of trained networks (Khorasani et al., 2026) and studied EU estimators qualitatively (Jiménez et al., 2026). To our knowledge, no prior work (i) grades uncertainty measures computed from the same ensemble against a computed EU target, location by location and with paired statistics, (ii) builds oracles for both reducible error and estimator spread on the same worlds and measures how far they agree, or (iii) compares credal functionals in the low-class regime typical of scientific classification.

Contributions.

1. **An oracle-referenced evaluation protocol for EU** (Section 3). We distinguish EU as the reducible error of a model from EU as the spread of its predictions when it is retrained on resampled data, and build an oracle for each by retraining the model many times in a simulated world where the true class probabilities are known. Measures are scored by how well they rank locations in the oracle's order, compared within each world. The protocol applies to any EU estimator.
2. **Credal uncertainty is more faithful than BMA when the ensemble is small, and only then** (Section 6.2). Summarising the credal set by its interval width (for each class, the gap between the largest and smallest probability any member assigns), two to five members rank locations more like the oracle than BMA does (by about 0.06 to 0.07 in rank correlation, significantly); for our main member model the advantage disappears at roughly ten members and reverses at forty, and the small-ensemble lead reproduces on real geochemical data.
3. **How the credal set is summarised matters more than whether one is used** (Section 6.1). The entropy interval used by Wang et al. (2025a), the difference between the highest and lowest entropy of any distribution in the set, shows no agreement with the oracle in two-class problems, while the width of the same set does; it improves as the number of classes grows. It depends on where the set lies as well as how wide it is, so it partly tracks the

confidence of the average prediction, which reflects label noise, rather than member disagreement, which more data would resolve.

4. **The advantage depends on the number of members, not the amount of data** (Section 6.3). Within the range tested, shrinking the training set from 500 to 100 samples leaves the credal advantage unchanged.

Outline. Section 2 reviews related work; Section 3 defines the estimands, oracles and grading; Section 4 the uncertainty measures; Sections 5 and 6 the setup and results, including real geochemistry; Section 7 guidance and limitations.

2 Related work

Ensemble-based EU and credal representations. Credal approaches keep a set of distributions rather than averaging ensemble members. Shaker & Hüllermeier (2021) compare Bayesian and credal ensemble inference; Wang et al. (2025a) introduce the post-hoc credal wrapper studied here; related work learns interval-valued outputs (Wang et al., 2024; 2025b), distils credal ensembles (Wang et al., 2026a) or uses credal priors (Caprio et al., 2024) (survey: Wang et al., 2026b). Evidential methods offer a different second-order representation (Sensoy et al., 2018). Random-set neural networks (Manchingal et al., 2023) predict a belief function over sets of classes, which induces a credal set whose width can serve as an EU measure (Manchingal, 2025); their outputs could be graded with the same protocol. None of our results depend on the specific wrapper: the protocol grades any per-input EU score, and the functional comparison applies to any method that outputs per-class probability intervals, including credal deep ensembles, interval networks and random-set networks.

Functionals for credal sets. A credal set is reduced to a scalar by a functional: non-specificity (Abellán & Moral, 2000), upper and lower entropy (Abellán & Moral, 2003), volume, or distance-based measures (Sale et al., 2024). Sale et al. (2023) show that volume is a sensible EU measure for binary classification but less so for many classes; our entropy-interval result runs the other way.

Evaluating EU. Analytical work has shown that common EU measures can violate desirable properties (Wimmer et al., 2023; Schweighofer et al., 2025) and that EU learned through loss minimisation need not be faithful (Bengs et al., 2022). Paplhám et al. (2026) argue that the right target is the expected regret of the deployed predictor. Khorasani et al. (2026) use class-conditional normalizing flows as exact-posterior oracles on images to measure the aleatoric floor and the reducible error of trained networks, and Jiménez et al. (2026) use simulated data-generating processes to study EU estimators qualitatively. We use such oracles instead to grade EU estimators.

EU in the geosciences. Mineral prospectivity mapping predicts where deposits are likely from sparse, expensive observations: drill holes, geochemical assays of soil and rock, and geophysical surveys. Datasets typically hold a few hundred to a few thousand samples, clustered around past exploration targets, and labels are few: deposit versus barren, or a handful of rock types. Machine learning, including Bayesian deep learning, is widely used for these tasks (Rodriguez-Galiano et al., 2018; Yuan et al., 2026; Huang et al., 2022; Liu, 2025). Because each hole is costly, exploration is increasingly framed as sequential decision-making under epistemic uncertainty, where the model's EU decides the next hole (Mern et al., 2024). In this literature EU quality is assessed only with proxies, and spatial autocorrelation is often handled by random splits, which inflate performance (Roberts et al., 2017); we address the first with the oracle protocol and measure the second on real data.

3 Oracle-referenced evaluation of epistemic uncertainty

This section defines what we measure and how we compute it. Figure 1 gives an overview, and Algorithm 1 in Section B gives the full procedure as pseudocode.

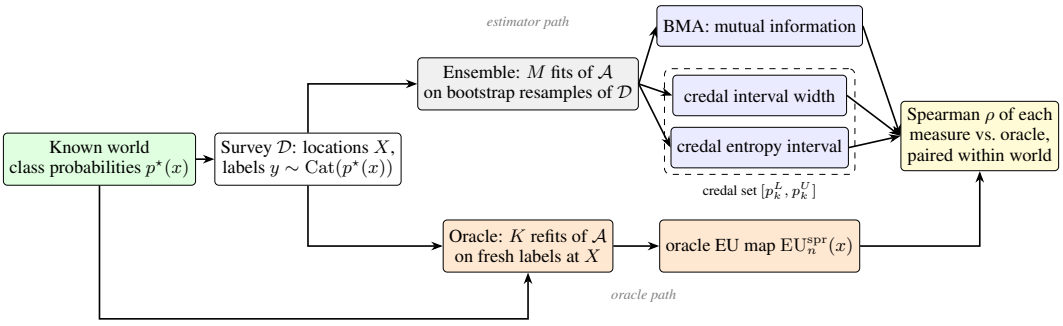

Figure 1: The evaluation protocol. On the estimator path, an ensemble is fitted on bootstrap resamples of a survey drawn from a known world, and its predictions are summarised by each uncertainty measure. On the oracle path, the same algorithm $\mathcal{A}$ is refitted on fresh labels at the same locations, giving the target EU map. Each measure is scored by its Spearman correlation with the oracle map, compared within each world.

3.1 Setting and notation

We consider classification with C classes. A *dataset* $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ consists of inputs $x_i \in \mathcal{X}$ and labels $y_i \in \{1, \dots, C\}$. In the geological setting an input is a location, a label is the rock type or deposit status found by drilling there, and a dataset is a drilling *survey*; we use the two words interchangeably. We call the set of locations $X = \{x_1, \dots, x_n\}$ the *design*, as distinct from the labels observed there. Each location has a true class-probability vector $p^\star(x) \in \Delta^{C-1}$, where $\Delta^{C-1} = \{p \in [0,1]^C : \sum_k p_k = 1\}$ is the probability simplex, and a drilled location returns a random label $y \sim \mathrm{Cat}(p^\star(x))$. We write $\mathcal{P}$ for the process that generates datasets in this way.

A *learning algorithm* $\mathcal{A}$ is a training procedure (model class, loss and optimiser) that maps a dataset to a fitted *model* $f = \mathcal{A}(\mathcal{D})$, which returns a predicted probability vector $f(x) \in \Delta^{C-1}$ at any location. A *refit* runs the same algorithm again on a different dataset. An *ensemble* is a set of M models $f_1, \dots, f_M$ fitted by $\mathcal{A}$ on bootstrap resamples of $\mathcal{D}$; its *members* predict $p_m(x) = f_m(x)$. An *uncertainty measure* maps the member predictions $p_1(x), \dots, p_M(x)$ to a single EU score. $H(p) = -\sum_k p_k \log_2 p_k$ is the Shannon entropy in bits.

3.2 Two definitions of epistemic uncertainty

Neither definition below is new. What we add is a way to compute both, so that estimators can be graded against them (Section 3.4).

Definition 1 (EU as reducible error)**.** $\mathrm{EU}_n^{\mathrm{red}}(x) = \mathbb{E}_{\mathcal{D}\sim\mathcal{P}^n}\left[\|f_n(x) - p^\star(x)\|^2\right]$, *where* $f_n = \mathcal{A}(\mathcal{D})$.

The expected squared (Brier) loss of a prediction $f(x)$ against a random label splits into two parts: $\mathbb{E}_y\|f(x) - e_y\|^2 = \|f(x) - p^\star(x)\|^2 + \left(1 - \|p^\star(x)\|^2\right)$, where e_y is the one-hot vector of y. The second term is the *aleatoric floor*: the loss that even the Bayes-optimal predictor $p^\star$ incurs because labels are random, which no amount of data can remove (Khorasani et al., 2026). Under log loss the floor is $H(p^\star(x))$. The first term is the part due to training on finite data; averaging it over datasets gives Definition 1, the excess risk used by Lahlou et al. (2023) and Paplhám et al. (2026), which vanishes as n grows for a consistent algorithm. We use squared error rather than log loss for Definition 1 because its excess over the floor has the closed form $\|f(x) - p^\star(x)\|^2$ and splits into bias and variance; the two oracles are compared by rank, so the difference in scale from Definition 2 does not matter. For example, next to dense drilling any model predicts close to $p^\star(x)$, so reducible error is small even where the label is a coin flip; far from any hole, predictions depend heavily on which survey was drawn, and reducible error is large.

Definition 2 (EU as estimator spread)**.** *Let* $q_1(x), \dots, q_K(x) \in \Delta^{C-1}$ *be the predictions at* x *of* K *refits of* $\mathcal{A}$ *on datasets with the same locations and fresh labels drawn from* $p^\star$*, and* $\bar{q}(x) = \frac{1}{K}\sum_k q_k(x)$ *their average. Then* $\mathrm{EU}_n^{\mathrm{spr}}(x) = H(\bar{q}(x)) - \frac{1}{K}\sum_{k=1}^K H(q_k(x))$.

This is the mutual-information decomposition used by BALD (Houlsby et al., 2011; Depeweg et al., 2018), applied to refits instead of posterior samples. The average $\bar{q}$ is the prediction obtained by pooling the refits. Definition 2 is the entropy of that pooled prediction minus the average entropy of the individual refits, so it is zero when all refits agree and grows as they disagree. For example, two refits predicting $(0.9, 0.1)$ and $(0.1, 0.9)$ pool to $(0.5, 0.5)$, and $\mathrm{EU}_n^{\mathrm{spr}} = 1 - 0.47 = 0.53$ bits. Two refits that both predict $(0.5, 0.5)$ give $\mathrm{EU}_n^{\mathrm{spr}} = 0$: each is uncertain, but that uncertainty is aleatoric and more data would not remove it.

When the two disagree. Reducible error splits into squared bias and variance (Geman et al., 1992): $\mathrm{EU}_n^{\mathrm{red}}(x) = \|\mathbb{E}_{\mathcal{D}} f_n(x) - p^\star(x)\|^2 + \mathbb{E}_{\mathcal{D}}\|f_n(x) - \mathbb{E}_{\mathcal{D}} f_n(x)\|^2$. Estimator spread reflects only the variance part. The two therefore agree wherever bias is small, and diverge under *shared bias*: every refit makes the same error, so spread is near zero while reducible error is large. This is why variance-based estimators can miss part of epistemic uncertainty (Lahlou et al., 2023; Jiménez et al., 2026). A model class too smooth to represent a narrow vein, for instance, smooths it out identically in every refit. Figure 2 (bottom left) shows where shared bias occurs in one of our worlds.

Two properties of the targets. First, both are relative to a learning algorithm and a sample size (Paplhám et al., 2026): EU is a property of a particular model trained on finite data, not of the world, and two algorithms trained on the same survey can be uncertain in different places. Every oracle is therefore built by refitting the same algorithm whose ensemble is being evaluated. Second, both refer to repeated draws of data from $\mathcal{P}$, not to a posterior over parameters; were EU defined as posterior spread, BMA would match the target by construction, whereas this definition favours neither method.

3.3 A CONTROLLED GENERATIVE WORLD

To compute either definition we need a world in which $p^\star$ is known and new surveys can be drawn at will. We use our own construction, chosen to be simple and cheap to sample rather than geologically realistic. Locations lie in the unit square $\mathcal{X} = [0, 1]^2$. Each class c has a latent score $z_c(x) = \sum_{b=1}^{B} a_{cb} \exp\big(-\|x - \mu_{cb}\|^2 / 2\sigma_{cb}^2\big)$, a sum of B Gaussian bumps with random centres μ_{cb}, widths σ_{cb} and amplitudes a_{cb}. Each bump can be read as a body of one rock type. The class probabilities are $p^\star(x) = \mathrm{softmax}(\beta z(x))$, where the inverse temperature $\beta > 0$ controls how sharp the boundaries between bodies are: large β gives near-certain labels and a low aleatoric floor, small β gives blurred boundaries and noisy labels. Surveys are drawn around a few random hubs, mimicking the clustered coverage of real drilling campaigns, and each drilled location returns a random label from $p^\star$. Figure 2 shows one world; parameters are in Section C.

3.4 TWO ORACLES

An *oracle* here is a Monte Carlo estimate of one of the two definitions, made possible because the world is known. Unlike earlier oracles (Khorasani et al., 2026; Jiménez et al., 2026), we build one for each definition on the same worlds and use them to grade uncertainty measures.

Disagreement oracle (Definition 2). Keep the design X fixed, draw fresh labels $y_i^{(k)} \sim \mathrm{Cat}(p^\star(x_i))$ for $k = 1, \ldots, K$, refit $q_k = \mathcal{A}(X, y^{(k)})$ on each, and evaluate Definition 2 at every point of a query grid. In words: if we drilled the same holes again and the outcomes came out differently, how much would the model change? A bootstrap ensemble estimates exactly this by resampling one dataset; the oracle computes it with genuinely new data, so a measure that disagrees with it is failing at its own task.

Learning-curve oracle (Definition 1). Draw R independent surveys of size n (new locations and new labels), fit $\mathcal{A}$ on each, and average $\|f(x) - p^\star(x)\|^2$ over the fits at every grid point. Repeating this at $n_0, 2n_0, 4n_0, \ldots$ traces a learning curve, which lets us check that the estimate really decays with n.

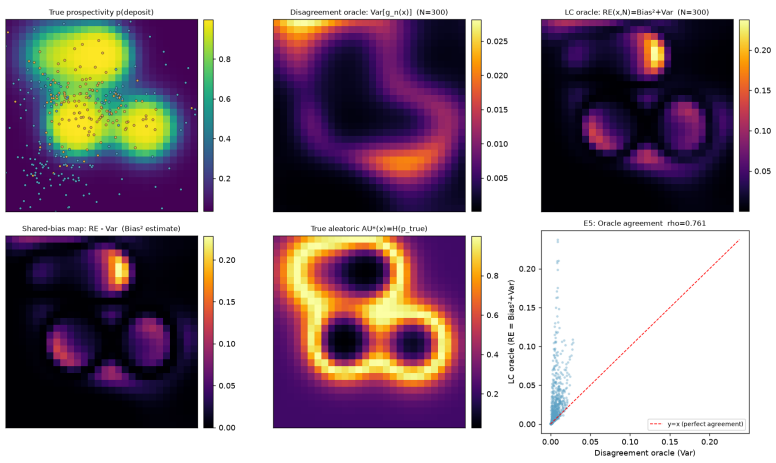

Figure 2: One synthetic world ($C = 2$, $n = 300$). Top row: true probability of a deposit $p^\star(x)$ with the drilled locations (left); disagreement oracle, here in its variance form (centre); learning-curve oracle, reducible error = bias^2+ variance (right). Bottom row: their difference, which marks shared-bias locations (left); aleatoric uncertainty $H(p^\star(x))$, high along the boundaries between bodies (centre); the two oracles plotted against each other over the query grid, Spearman $\rho = 0.761$ (right).

The learning-curve oracle captures bias as well as variance, while the disagreement oracle captures only the spread. Where reducible error is high but refits agree lies shared bias, which no disagreement-based measure can detect; Section 6.4 reports how closely the two maps agree.

3.5 Grading an uncertainty measure

What matters is the ranking. In sequential acquisition, a form of active learning (Settles, 2009), the model is retrained after each drilling round, scores every candidate location, and the top-scoring one is drilled next (Mern et al., 2024); only the order of the scores matters. We therefore score a measure by the Spearman rank correlation ρ (Spearman, 1904) between its EU map and the oracle map over the G locations of the query grid. ρ is the ordinary correlation between the two sets of ranks: 1 means the same ordering, 0 no relation, and -1 the reverse ordering. We call ρ the *fidelity* of the measure.

Comparing measures within each world. Results are averaged over S independently generated worlds, indexed by s. Within world s, all measures are computed from the same ensemble and compared with the same oracle, so the fidelities $\rho_A^{(s)}$ and $\rho_B^{(s)}$ of two measures A and B (for example credal width and BMA) are *paired*: a world that is hard for one measure tends to be hard for the other. We therefore analyse the difference within each world, $\Delta\rho^{(s)} = \rho_A^{(s)} - \rho_B^{(s)}$, which cancels the variation in difficulty between worlds, much as a paired t-test does. We report three numbers: the mean of $\Delta\rho^{(s)}$ over worlds; a 95% bootstrap interval, obtained by resampling the S differences with replacement 20,000 times and taking the 2.5th and 97.5th percentiles of the resampled means (Efron & Tibshirani, 1994); and the P-value of a Wilcoxon signed-rank test (Wilcoxon, 1945), which asks whether the differences are centred at zero without assuming they are normally distributed. P is the probability of a difference at least this large if the two measures were equally faithful (capital P avoids confusion with probability vectors p).

4 Uncertainty measures under test

Each member's prediction $p_m(x)$ is a probability vector over the C classes, i.e. a point in the simplex Δ^{C-1} (Section 3.1). All uncertainty measures below take the same set $\{p_1(x), \ldots, p_M(x)\}$, so any difference between them is due to the measure alone.

BMA. $\mathrm{EU}_{\mathrm{BMA}} = H(\bar{p}) - \frac{1}{M}\sum_m H(p_m)$ with $\bar{p} = \frac{1}{M}\sum_m p_m$: the members are averaged first and the uncertainty is read from the blend, as in Definition 2.

Credal set. Following Wang et al. (2025a), per-class bounds $p_k^L = \min_m p_{m,k}$ and $p_k^U = \max_m p_{m,k}$ define the credal set $\mathcal{P} = \{p \in \Delta^{C-1} : p_k^L \le p_k \le p_k^U \ \forall k\}$, the set of all distributions consistent with the range of member predictions. Any method that outputs lower and upper class probabilities induces the same kind of set, so the two functionals below apply equally to those methods. We summarise the set in two ways:

$$\mathrm{EU}_{\mathrm{width}} = \tfrac{1}{C}\textstyle\sum_{k=1}^{C}\left(p_k^U - p_k^L\right), \tag{1}$$

$$\mathrm{EU}_{\mathrm{ent}} = \overline{H}(\mathcal{P}) - \underline{H}(\mathcal{P}), \quad \overline{H} = \max_{p\in\mathcal{P}} H(p), \ \ \underline{H} = \min_{p\in\mathcal{P}} H(p). \tag{2}$$

The width depends only on how far apart the members are, so it is a pure disagreement measure, in line with its use as an EU measure in random-set networks (Manchingal, 2025). The entropy interval is the functional used by Wang et al. (2025a).

Lemma 1 (Upper entropy by water-filling)**.** *The distribution of maximum entropy in $\mathcal{P}$ has coordinates $p_k = \mathrm{median}(p_k^L, t, p_k^U)$, where the level t is the unique value for which $\sum_k p_k = 1$.*

This is a standard consequence of the KKT conditions (proof in Section A); algorithms for maximum entropy over probability intervals are also given by Abellán & Moral (2003). The level t has a simple reading, shown in Figure 4: maximum entropy makes the class probabilities as equal as the intervals allow. Every class takes the common value t unless its interval forces it higher or lower, and t is raised until the probabilities sum to one, like water poured into columns with floors p_k^L and ceilings p_k^U. Because $\overline{H}$ can be computed exactly this way for any number of classes, we can evaluate the entropy interval reliably from $C = 2$ up to $C = 20$ in Section 6.1. $\underline{H}$ is computed by a greedy construction over the vertices of $\mathcal{P}$, which is exact for $C = 2$ and approximate for $C > 2$.

Baseline. We also report MC-Dropout (Gal & Ghahramani, 2016): a single network with dropout kept active at prediction time, whose M stochastic forward passes are treated as ensemble members and summarised with the BMA measure. It gives a cheaper ensemble whose diversity comes from dropout masks rather than from resampling the data.

5 Experimental setup

Member model. The algorithm $\mathcal{A}$ fits random-Fourier-feature multinomial logistic models (Rahimi & Recht, 2007): $D = 128$ shared random features approximating an RBF kernel, followed by ℓ_2-regularised logistic regression. Ensemble diversity comes only from bootstrap resampling of the drilled data, as in a deep ensemble. We chose it for two reasons. The oracles need 51,200 refits, feasible only if a fit takes milliseconds; random-feature logistic regression approximates a Gaussian-process classifier at that cost (under 30 CPU-minutes in total, Section K). And the comparison depends only on member predictions, so the member model should matter little, which Section 6.4 tests with neural-network members and on real data.

Protocol. Each setting uses $S = 40$ independently generated worlds, a 28×28 query grid ($G = 784$) and $K = 80$ oracle refits. The oracle does not depend on M, so it is computed once per world and reused across ensemble sizes, which makes the pairing exact. The three sweeps are: **class count** $C \in \{2, 3, 5, 10, 20\}$ at $M = 5$, with $n = 60C$ so samples per class stay roughly constant (a fixed-n arm is in Section D); **number of members** $M \in \{2, 3, 5, 10, 20, 40\}$ at $n = 150$, $C = 2$; and **sample size**, retaining 100/80/60/40/20% of a survey of $n = 500$ at $M = 5$, $C = 2$.

Hypothesis for the class-count sweep. The width depends only on how far apart the members are. The entropy interval also depends on where the interval sits relative to the maximum-entropy point of the simplex, so two locations with identical member disagreement can receive different scores. This is undesirable: member disagreement is what more data would reduce, whereas the interval's position largely tracks the confidence of the average prediction, which is closer to aleatoric uncertainty. In binary classification entropy is a single symmetric arc in one coordinate, and this location effect should be at its strongest. With more classes, entropy depends on the whole probability vector

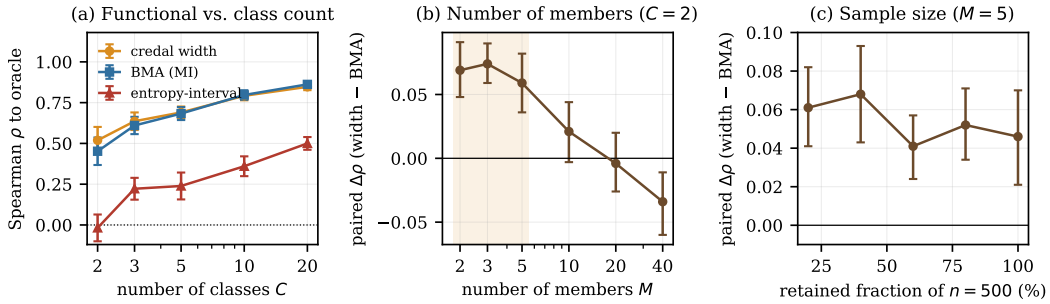

Figure 3: Fidelity to the disagreement oracle (40 worlds per point). (a) Fidelity of each measure versus class count at fixed samples per class, $M = 5$; bars are 95% intervals on the mean. (b, c) Paired per-world gap between credal width and BMA with bootstrap 95% intervals; shading marks $M \leq 5$.

rather than on one coordinate's distance from $1/2$, and the effect should dilute. To our knowledge this has not been noted before for the entropy interval (cf. Wimmer et al., 2023).

6 Results

6.1 Functional choice depends on class count

Figure 3a shows three results (full numbers in Table 2, Section D).

The entropy interval carries no epistemic signal in binary classification. At $C = 2$ its fidelity is -0.018 (and 0.007 in the fixed-n arm), indistinguishable from zero, while the interval width of the *same* credal set reaches 0.519. The gap between the two summaries, $\Delta\rho = +0.54$, is roughly eight times the gap between the credal width and BMA ($+0.067$). This is the location effect hypothesised in Section 5: in the binary case, where the interval sits in the simplex swamps how wide it is.

The effect weakens as classes are added. The entropy interval's fidelity rises monotonically from -0.018 to 0.500 between $C = 2$ and $C = 20$, and its gap to the width falls from $+0.537$ to $+0.348$. This is consistent with its good performance on the many-class image benchmarks on which credal wrappers were introduced (Wang et al., 2025a), where the binary regime was not the focus. At $C = 2$ the width equals the credal volume, so our binary result agrees with Sale et al. (2023), who find volume a good EU measure for binary problems but weaker with more classes, the opposite of the entropy interval. Credal functionals therefore do not share one regime of validity, and should be chosen for the class count at hand.

The credal advantage over BMA is also largest at small C: $+0.067$ at $C = 2$ ($P < 10^{-7}$), not significant by $C = 5$, and slightly reversed at $C = 20$, in both arms (Section D).

6.2 The credal advantage is a small-ensemble effect

Figure 3b shows the sweep over the number of members (full results in Table 4, Section E). The credal width leads BMA by $\Delta\rho \approx +0.06$ to $+0.07$ at every $M \leq 5$ ($P < 10^{-4}$), and wins in 34 to 35 of 40 worlds (Section E). This is the regime targeted by Wang et al. (2025a), and it confirms their small-ensemble claim against an oracle reference. The lead is no longer significant at $M = 10$ ($P = 0.084$) or $M = 20$ ($P = 0.868$; width wins exactly 20 of 40 worlds), and at $M = 40$ BMA is ahead ($\Delta\rho = -0.034$, $P = 0.010$). The crossover lies between roughly 10 and 20 members. MC-Dropout trails at every M, and the entropy interval stays near zero.

6.3 The advantage does not grow as data thin

Credal representations are usually motivated by scarce evidence, so one might expect the advantage to grow as the training set shrinks. Figure 3c shows otherwise. The width leads BMA significantly

at every retained fraction ($P < 10^{-3}$ throughout; Section F), but the gap fluctuates between $+0.041$ and $+0.068$ with no trend as n falls from 500 to 100.

If the credal set helped because data are scarce, its advantage would grow as n falls; it does not, yet at fixed n it shrinks steadily as members are added (Section 6.2). The advantage is therefore controlled by M, the number of samples describing the model's uncertainty, not by the size of the training set. One untested explanation is that the plug-in mutual-information estimate is noisy at small M, while the range is robust at small M but dominated by extreme members at large M. We make no claim beyond $n = 500$ (Section G).

6.4 Robustness

Do the two estimands agree? On a fixed design of $n = 300$, the disagreement and learning-curve oracles have rank correlation $\rho = 0.761$ across the query grid (Figure 2, bottom right), so rankings graded against spread largely carry over to reducible error; shared bias concentrates at the edges of dense drilling, where models extrapolate (Figure 2; Section H).

Member model. Repeating the sweep over the number of members with bootstrap MLPs trained on raw coordinates in place of the random-feature members (Section I) preserves the sign of the small-ensemble gap ($\Delta\rho = +0.040$, $P = 0.015$ at $M = 2$), although the MLP gaps at $M = 3$ and 5 are not individually significant and the MLP run shows no reversal up to $M = 40$. The small-M ranking is therefore not an artefact of the random feature map, but the crossover point depends on the member model.

Real geochemistry. To check that the result is not a property of the simulator, we take the densest 2° cell of the South Australian geochemical database (Geological Survey of South Australia, 2023) (1,218 samples, six lithological classes, nine centred-log-ratio element features), treat it as the population, draw 40 sparse surveys of $n = 150$, and compute the oracle from $K = 80$ refits on each (Section J). Table 8 (Section J) shows the same pattern as Section 6.2: the width leads at $M \leq 5$ ($P \leq 0.002$) and BMA leads from $M = 10$. The margin is smaller ($+0.01$ to $+0.03$), as expected at $C = 6$ (Section 6.1).

On the full dataset under spatial block cross-validation, credal and BMA measures of the same backbones are within fold noise in balanced accuracy, and random splits inflate it by about 0.20 for every method (Section J).

7 Discussion and limitations

Practical guidance. With fewer than about ten members and few classes, read the credal set through its interval width, which is also about four orders of magnitude cheaper than the entropy interval (Section K); avoid the entropy interval in binary problems; with large ensembles or many classes, BMA is at least as good.

Limitations. (i) The disagreement oracle conditions on the survey design; a design-marginal target would be broader. (ii) The synthetic world is one family of smooth fields, which the real-data oracle only partly mitigates. (iii) The crossover point appears to depend on the member model. (iv) $\underline{H}$ is exact only at $C = 2$. (v) The credal versus BMA effects are small ($\Delta\rho \leq 0.07$) though consistent; the functional effect ($\Delta\rho \approx 0.5$) is large. (vi) EU is model-specific, so we rank measures, not algorithms. Next, the protocol can compare further credal functionals (volume, non-specificity) against a common oracle.

8 Conclusion

We introduced a protocol that grades EU estimators against computed oracles for reducible error and estimator spread, rather than proxies. It applies to any EU estimator whose data-generating process can be simulated. It shows that the credal interval width beats BMA with few ensemble members but not many, that the entropy interval carries no epistemic signal in two-class problems, and that the advantage tracks the number of members, not the amount of data.

AI USE STATEMENT

We used generative AI tools to help debug the code used for our experiments and to check the grammar and quality of language in the text. All numerical results were produced by the authors' code, and all AI-assisted code and text were reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses openly licensed geochemical survey data (Geological Survey of South Australia, 2023; United States Geological Survey, 2023) and synthetic data; no human subjects or personal data are involved. Better-calibrated exploration models could make drilling more targeted and so reduce ground disturbance, but they may also accelerate extraction whose costs fall on local communities; improved uncertainty estimates do not resolve that asymmetry, and we view such tools as informing rather than replacing expert and community decision-making.

REPRODUCIBILITY STATEMENT

The generative world, both oracles, all uncertainty measures and the scripts that regenerate every table and figure are provided as anonymised supplementary code. Algorithm 1 gives the full procedure, all hyperparameters are listed in Section C, the proof of Lemma 1 is in Section A, and real-data preprocessing, including the spatial cross-validation protocol and dataset identifiers, is described in Section J. Every experiment is seeded, and per-world results are released as CSV files.

REFERENCES

Joaquín Abellán and Serafín Moral. A non-specificity measure for convex sets of probability distributions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 8(3): 357–367, 2000.

Joaquín Abellán and Serafín Moral. Maximum of entropy for credal sets. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(5):587–597, 2003.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Advances in Neural Information Processing Systems*, 2022.

Michele Caprio, Souradeep Dutta, Kuk Jin Jang, Vivian Lin, Radoslav Ivanov, Oleg Sokolsky, and Insup Lee. Credal Bayesian deep learning. *Transactions on Machine Learning Research*, 2024.

Fabio Cuzzolin. *The Geometry of Uncertainty: The Geometry of Imprecise Probabilities*. Springer Nature, 2020.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pp. 1184–1193, 2018.

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

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Geological Survey of South Australia. Geochemical survey data repository (SARIG). `https://energymining.sa.gov.au/minerals/geoscience/geoscientific_data`, 2023.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Dapeng Huang, Renguang Zuo, and Jian Wang. Geochemical anomaly identification and uncertainty quantification using a Bayesian convolutional neural network model. *Applied Geochemistry*, 146: 105450, 2022.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Sebastián Jiménez, Mira Jürgens, and Willem Waegeman. Position: Epistemic uncertainty estimation methods are fundamentally incomplete. In *International Conference on Machine Learning*, 2026.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on Bayesian neural networks: A tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arian Khorasani, Nathaniel Chen, Yug D. Oswal, Akshat Santhana Gopalan, Egemen Kolemen, and Ravid Shwartz-Ziv. Beyond the loss curve: Scaling laws, active learning, and the limits of learning from exact posteriors. *arXiv preprint arXiv:2602.00315*, 2026.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yue Liu. Quantifying uncertainty of mineral prediction using a novel Bayesian deep learning framework. *Artificial Intelligence in Geosciences*, pp. 100164, 2025.

Shireen Kudukkil Manchingal. *Epistemic Deep Learning: Enabling Machine Learning Models to 'Know When They Do Not Know'*. PhD thesis, Oxford Brookes University, 2025. arXiv:2510.22261.

Shireen Kudukkil Manchingal, Muhammad Mubashar, Kaizheng Wang, Keivan Shariatmadar, and Fabio Cuzzolin. Random-set convolutional neural network (RS-CNN) for epistemic deep learning. *arXiv preprint arXiv:2307.05772*, 2023.

John Mern et al. Intelligent prospector v2.0: Exploration drill planning under epistemic model uncertainty. *arXiv preprint arXiv:2410.10610*, 2024.

Jakub Paplhám, Willem Waegeman, Eyke Hüllermeier, and Vojtěch Franc. Evaluating epistemic uncertainty: Beyond OOD detection and active learning. *arXiv preprint arXiv:2607.14817*, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40 (8):913–929, 2017.

Victor Rodriguez-Galiano, Manuel Sanchez-Castillo, Mario Chica-Olmo, and M. Chica-Rivas. Classification of lithostratigraphic and alteration units from drillhole lithogeochemical data using machine learning. *Ore Geology Reviews*, 89:78–90, 2018.

Yusuf Sale, Michele Caprio, and Eyke Hüllermeier. Is the volume of a credal set a good measure for epistemic uncertainty? In *Conference on Uncertainty in Artificial Intelligence*, volume 216 of *PMLR*, pp. 1795–1804, 2023.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *International Conference on Machine Learning*, 2024.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. On information-theoretic measures of predictive uncertainty. In *Conference on Uncertainty in Artificial Intelligence*, 2025.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

Mohammad Hossein Shaker and Eyke Hüllermeier. Ensemble-based uncertainty quantification: Bayesian versus credal inference. *Proceedings of the Workshop on Computational Intelligence*, 25:63–79, 2021.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

United States Geological Survey. National geochemical survey database. `https://mrdata.usgs.gov/geochem/`, 2023.

Kaizheng Wang, Fabio Cuzzolin, Shireen Kudukkil Manchingal, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal deep ensembles for uncertainty quantification. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. Credal wrapper of model averaging for uncertainty estimation in classification. In *International Conference on Learning Representations*, 2025a.

Kaizheng Wang, Keivan Shariatmadar, Shireen Kudukkil Manchingal, Fabio Cuzzolin, David Moens, and Hans Hallez. CreINNs: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, 185:107198, 2025b.

Kaizheng Wang, Fabio Cuzzolin, David Moens, and Hans Hallez. Credal ensemble distillation for uncertainty quantification. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 26319–26327, 2026a.

Kaizheng Wang, Fabio Cuzzolin, Keivan Shariatmadar, David Moens, and Hans Hallez. A review of uncertainty representation and quantification in neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2476–2495, 2026b.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Manbin Yuan et al. A systematic review of deep learning methods for mineral exploration using multisource geoscience data (2018–2025). *International Journal of Applied Earth Observation and Geoinformation*, 149:105254, 2026.

Guide to the appendix. Section A proves Lemma 1; Section B gives the protocol as pseudocode; Section C lists all hyperparameters; Section D reports the fixed-n class-count sweep; Section E gives per-world result distributions; Section F the full sample-size results; Section G a repeat at a larger design; Section H details of the oracle-agreement experiment; Section I the member-model ablation; Section J real-data preprocessing and spatial cross-validation; and Section K computational cost.

A Proof of Lemma 1

$H(p) = -\sum_k p_k \log_2 p_k$ is separable, strictly concave and symmetric in its coordinates, and $\mathcal{P}$ is the intersection of the box $\prod_k [p_k^L, p_k^U]$ with the hyperplane $\sum_k p_k = 1$, assumed non-empty. The maximiser is therefore unique. The KKT conditions read $-\log_2 p_k - 1/\ln 2 = \lambda + \nu_k - \eta_k$ with $\nu_k, \eta_k \geq 0$ the multipliers for $p_k \leq p_k^U$ and $p_k \geq p_k^L$ and complementary slackness. For every coordinate strictly inside its box, $\nu_k = \eta_k = 0$, so all such coordinates share the common value $t = 2^{-\lambda - 1/\ln 2}$. A coordinate at its upper bound has $\nu_k \geq 0$, hence $p_k^U \leq t$; a coordinate at its lower bound has $p_k^L \geq t$. In all cases $p_k = \mathrm{median}(p_k^L, t, p_k^U)$. The map $t \mapsto \sum_k \mathrm{median}(p_k^L, t, p_k^U)$ is continuous and non-decreasing, ranging from $\sum_k p_k^L \leq 1$ to $\sum_k p_k^U \geq 1$, so a level satisfying the constraint exists and yields the unique maximiser; it is located by bisection (80 iterations). □

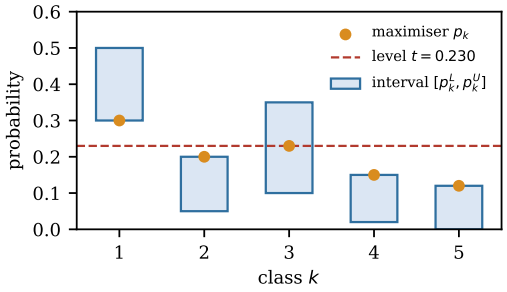

Figure 4: Water-filling for the upper entropy with $C = 5$. Classes 2, 4 and 5 are held at their upper bounds, class 1 at its lower bound, and class 3 sits at the level t.

B Protocol pseudocode

Algorithm 1 Oracle-referenced grading of EU measures

Require: learning algorithm $\mathcal{A}$; survey size n; ensemble size M; oracle refits K; query grid G; uncertainty measures $U_1, \ldots, U_J$; number of worlds S

for $s = 1, \ldots, S$ **do**
  Draw a world $p^\star$ and a design $X = \{x_1, \ldots, x_n\}$; draw labels $y_i \sim \mathrm{Cat}(p^\star(x_i))$
  Estimator path
  for $m = 1, \ldots, M$ **do**
    Resample (X, y) with replacement to get $\mathcal{D}_m$; fit $f_m \leftarrow \mathcal{A}(\mathcal{D}_m)$; set $p_m(x) \leftarrow f_m(x)$ for $x \in G$
  end for
  For each measure j and each $x \in G$, compute $U_j(x)$ from $p_1(x), \ldots, p_M(x)$
  Oracle path
  for $k = 1, \ldots, K$ **do**
    Draw fresh labels $y_i^{(k)} \sim \mathrm{Cat}(p^\star(x_i))$ at the same design; refit $q_k \leftarrow \mathcal{A}(X, y^{(k)})$
  end for
  $\mathrm{EU}^{\mathrm{spr}}(x) \leftarrow H(\bar{q}(x)) - \frac{1}{K}\sum_k H(q_k(x))$ for $x \in G$, where $\bar{q} = \frac{1}{K}\sum_k q_k$
  $\rho_j^{(s)} \leftarrow \mathrm{Spearman}\big(U_j, \mathrm{EU}^{\mathrm{spr}}\big)$ over G, for each j
end for
For each pair (A, B) of measures: $\Delta\rho^{(s)} \leftarrow \rho_A^{(s)} - \rho_B^{(s)}$; report the mean, a bootstrap 95% interval and a Wilcoxon signed-rank P-value over $s = 1, \ldots, S$

C SETTINGS

Table 1: Hyperparameters used for all simulation results unless stated otherwise.

Component	Parameter	Value
World	domain; bodies per class B	$[0,1]^2$; 3
	centres; widths; amplitudes	$\mathcal{U}[0.05, 0.95]^2$; $\mathcal{U}[0.10, 0.28]$; $\mathcal{U}[0.8, 2.0]$
	inverse temperature β	2.5
Survey	hubs; scatter about hub	4 drawn from $\mathcal{U}[0.12, 0.88]^2$; $\mathcal{N}(0, 0.11^2)$, clipped
Member model	random features D; bandwidth γ	128 (shared across members); 9.0, $W \sim \mathcal{N}(0, 2\gamma)$
	inverse regularisation; diversity	1.0 (ℓ_2); bootstrap of size n
Oracle	refits K; query grid	80; 28×28 on $[0.02, 0.98]^2$
Credal	$\overline{H}$; $\underline{H}$	bisection, 80 iterations; greedy vertex construction
Inference	worlds per cell; test	40; paired two-sided Wilcoxon signed-rank

D CLASS-COUNT SWEEP: FULL RESULTS AND FIXED-n ARM

Table 2: Class-count sweep at fixed samples per class ($n = 60C$, $M = 5$, 40 worlds). Mean Spearman fidelity to the oracle, and paired gaps with bootstrap 95% intervals and Wilcoxon P-values.

		Fidelity ρ			width − entropy interval		width − BMA	
C	n	width	BMA	ent. int.	$\Delta\rho$ [95% CI]	P	$\Delta\rho$ [95% CI]	P
2	120	**0.519**	0.452	−0.018	+0.537 [+0.465, +0.610]	$< 10^{-11}$	+0.067 [+0.049, +0.084]	$< 10^{-7}$
3	180	**0.636**	0.610	0.222	+0.414 [+0.344, +0.481]	$< 10^{-10}$	+0.026 [+0.012, +0.041]	$< 10^{-3}$
5	300	**0.690**	0.683	0.239	+0.451 [+0.370, +0.530]	$< 10^{-10}$	+0.007 [−0.008, +0.022]	0.368
10	600	0.792	**0.797**	0.360	+0.432 [+0.377, +0.489]	$< 10^{-11}$	−0.005 [−0.015, +0.004]	0.397
20	1200	0.847	**0.862**	0.500	+0.348 [+0.310, +0.385]	$< 10^{-11}$	−0.015 [−0.026, −0.005]	0.003

In this arm $n = 400$ for every C, so samples per class fall as C grows (to a minimum class size of 1.3 on average at $C = 20$). The entropy interval is again uninformative at $C = 2$, and the width − BMA gap again favours the width only at small C. The non-monotone recovery of the entropy interval here reflects the collapse of per-class data at large C, which is why the fixed-samples-per-class arm is the controlled comparison.

Table 3: Fixed $n = 400$, $M = 5$, 40 worlds. Mean fidelity (s.d.) and paired gaps.

C	width	BMA	ent. int.	$\Delta\rho$ (w − ent)	P	$\Delta\rho$ (w − BMA) [95% CI]	P
2	0.634 (0.167)	0.568 (0.207)	0.007 (0.247)	+0.627	$< 10^{-11}$	+0.067 [+0.043, +0.093]	$< 10^{-6}$
3	0.719 (0.127)	0.703 (0.131)	0.255 (0.186)	+0.464	$< 10^{-11}$	+0.016 [+0.001, +0.033]	0.128
5	0.765 (0.103)	0.758 (0.106)	0.322 (0.194)	+0.443	$< 10^{-11}$	+0.007 [−0.003, +0.017]	0.320
10	0.750 (0.140)	0.755 (0.138)	0.295 (0.191)	+0.454	$< 10^{-11}$	−0.006 [−0.017, +0.005]	0.320
20	0.775 (0.109)	0.792 (0.116)	0.233 (0.205)	+0.542	$< 10^{-11}$	−0.016 [−0.029, −0.004]	0.024

E NUMBER OF MEMBERS: FULL RESULTS AND PER-WORLD DISTRIBUTIONS

Table 4: Sweep over the number of members M ($n = 150$, $C = 2$, 40 worlds): mean fidelity and paired width − BMA gap.

M	width	BMA	ent. int.	MC-Dropout	$\Delta\rho$ (width − BMA) [95% CI]	P
2	**0.330**	0.261	−0.000	0.073	+0.069 [+0.048, +0.091]	$< 10^{-6}$
3	**0.421**	0.347	−0.036	0.089	+0.074 [+0.059, +0.090]	$< 10^{-9}$
5	**0.475**	0.416	−0.052	0.253	+0.059 [+0.036, +0.082]	$< 10^{-4}$
10	**0.589**	0.568	−0.024	0.345	+0.021 [−0.003, +0.044]	0.084
20	0.671	**0.675**	0.054	0.380	−0.004 [−0.026, +0.020]	0.868
40	0.684	**0.719**	0.007	0.459	−0.034 [−0.060, −0.011]	0.010

Table 5: Per-world distribution of $\Delta\rho = \rho_{\text{width}} - \rho_{\text{BMA}}$. Left: sweep over the number of members. Right: class-count sweep, fixed n.

M	wins	median	min	max
2	35/40	+0.064	−0.073	+0.245
3	35/40	+0.080	−0.033	+0.194
5	34/40	+0.062	−0.121	+0.302
10	27/40	+0.029	−0.160	+0.174
20	20/40	+0.001	−0.153	+0.172
40	12/40	−0.027	−0.277	+0.088

C	wins	median	min	max
2	31/40	+0.048	−0.040	+0.304
3	24/40	+0.004	−0.071	+0.157
5	19/40	−0.001	−0.047	+0.099
10	16/40	−0.002	−0.143	+0.076
20	10/40	−0.017	−0.115	+0.058

F SAMPLE-SIZE SWEEP

Table 6: Retained fraction of a survey of $n = 500$ ($M = 5$, $C = 2$, 40 worlds).

Retained	n	width	BMA	ent. int.	MC-Dropout	$\Delta\rho$ (width − BMA) [95% CI]	P
100%	500	**0.651**	0.606	0.109	0.280	+0.046 [+0.021, +0.070]	$< 10^{-3}$
80%	400	**0.622**	0.570	0.089	0.308	+0.052 [+0.034, +0.071]	$< 10^{-5}$
60%	300	**0.629**	0.589	0.082	0.368	+0.041 [+0.024, +0.057]	$< 10^{-4}$
40%	200	**0.460**	0.393	−0.054	0.190	+0.068 [+0.043, +0.093]	$< 10^{-5}$
20%	100	**0.516**	0.455	−0.045	0.198	+0.061 [+0.041, +0.082]	$< 10^{-6}$

G NUMBER OF MEMBERS AT TWO DESIGN SIZES

We repeated the sweep over the number of members at $n \in \{150, 600\}$ ($C = 2$). At $M = 5$, the setting emphasised by Wang et al. (2025a), the width leads BMA at $n = 150$ ($\Delta\rho = +0.039$, $P = 0.001$); at $n = 600$ the gap is not significant ($\Delta\rho = -0.009$, $P = 0.261$). In both cases BMA overtakes the width between $M \approx 10$ and $M \approx 20$. The $n = 600$ result is why Section 6.3 restricts its claim to the range $n \in [100, 500]$.

H ORACLE AGREEMENT

On a fixed design of $n = 300$ drills, the disagreement oracle uses $K = 100$ label redraws and the learning-curve oracle uses $K = 60$ sub-surveys to estimate the reducible error $\text{RE}(x, n) = \mathbb{E}[(g_n(x) - p^\star(x))^2] = \text{Bias}^2 + \text{Var}$. In this experiment the disagreement oracle is measured by the prediction variance rather than the mutual-information form of Definition 2. The two oracles have Spearman correlation 0.761 over the grid. The shared-bias map $\hat{B}^2 = \max(\text{RE} - \widehat{\text{Var}}, 0)$ is concentrated at the periphery of dense drilling; we report it qualitatively, since a pointwise positive estimate at a given location need not be distinguishable from estimation noise.

I MEMBER-MODEL ABLATION

Table 7: Paired gap $\Delta\rho = \rho_{\text{width}} - \rho_{\text{BMA}}$ for two member models ($n = 150$, $C = 2$, 40 worlds). The random-feature model here uses $D = 256$ features, lengthscale 0.25; the MLP has 32 and 16 ReLU units, $\ell_2 = 0.05$, on raw coordinates.

	Random-feature logistic		MLP	
M	$\Delta\rho$	P	$\Delta\rho$	P
2	$+0.073$	< 0.001	$+0.040$	0.015
3	$+0.061$	< 0.001	$+0.027$	0.090
5	$+0.028$	0.064	$+0.041$	0.070
10	-0.013	0.277	$+0.055$	0.030
20	-0.026	0.015	$+0.041$	0.044
40	-0.043	0.001	$+0.073$	0.004

J REAL-DATA PREPROCESSING AND DEPLOYMENT RESULTS

Data. Drill-hole assays from the South Australian Resources Information Gateway (Geological Survey of South Australia, 2023), released under the South Australian Government open-data licence. Values below detection are imputed at half the detection limit; nine elements (Cu, Si, Al, Fe, Mg, Ca, Na, K, Ti) are retained on coverage; features are centred-log-ratio transformed and standardised on training folds only. The processed data contain 9,473 samples in six lithological classes (chance $1/6$).

Table 8: Resampling oracle on real geochemistry (dense SARIG sub-region, $N_{\text{pop}} = 1{,}218$, $C = 6$, $n = 150$, $K = 80$, 40 sub-surveys): mean fidelity and paired gap $\Delta\rho$ (width $-$ BMA).

M	2	3	5	10	20	40
BMA	0.409	0.528	0.653	**0.763**	**0.841**	**0.902**
width	**0.441**	**0.555**	**0.664**	0.752	0.803	0.843
$\Delta\rho$	$+0.032$	$+0.027$	$+0.011$	-0.012	-0.038	-0.059
P	< 0.001	< 0.001	0.002	< 0.001	< 0.001	< 0.001

Resampling oracle. The densest 2° cell (136–138°E, 34–32°S) serves as the population ($N_{\text{pop}} = 1{,}218$). For each of 40 sparse surveys of $n = 150$, $K = 80$ bootstrap refits give the oracle on the remaining $\approx 1{,}068$ locations, and an RFF-logistic ensemble ($D = 256$, lengthscale 1.0) with up to 40 members supplies the measures.

Spatial versus random cross-validation. Spatial blocks are formed by k-means on site coordinates and assigned to five folds (Roberts et al., 2017).

Table 9: Balanced accuracy under spatial block and random 5-fold CV ($C = 6$). Credal rows share their backbone with the BMA row above them; only the measure differs.

Method	Backbone	Params	Spatial	Random	Leak
Single NN	MLP 9–128–128–6	18,566	0.398	0.620	$+0.223$
Temp.-scaled NN	MLP $+T$	18,567	0.398	0.620	$+0.223$
Deep ensemble (BMA)	$5\times$ MLP	92,830	0.399	0.615	$+0.216$
Credal (ensemble)	same	92,830	0.401	0.610	$+0.209$
MC-Dropout (BMA)	MLP, $p = 0.3$	18,566	0.426	0.619	$+0.193$
Credal (MC-Dropout)	same	18,566	0.423	0.614	$+0.191$
Variational BNN (BMA)	DenseFlipout	37,132	0.431	0.596	$+0.165$
Credal (BNN)	same	37,132	0.435	0.592	$+0.157$

K COMPUTATIONAL COST

Table 10: Wall-clock cost per world at $n = 400$ over 784 query points, single CPU core.

Component	$C = 2$	$C = 5$	$C = 20$
Oracle, $K = 80$ refits	0.46 s	0.65 s	1.33 s
Ensemble, $M = 5$	24 ms	37 ms	65 ms
Measure: BMA	0.17 ms	0.26 ms	0.50 ms
Measure: credal width	0.03 ms	0.04 ms	0.14 ms
Measure: entropy interval	247 ms	320 ms	510 ms

The completed simulation programme (640 worlds, 51,200 oracle refits, about 4,000 ensemble fits) runs in under 30 minutes on one CPU core with no accelerator.