REALITrees: Rashomon Ensemble Active Learning for Interpretable Trees

Simon D. Nguyen[1] **Hayden McTavish**[3] **Kentaro Hoffman**[1] **Tyler H. McCormick**[1,2] **Cynthia Rudin**[3]

[1]Department of Statistics, University of Washington
[2]Department of Sociology, University of Washington
[3]Department of Computer Science, Duke University

Abstract

Active learning reduces labeling costs by querying samples to maximize information gain. A dominant framework, Query-by-Committee, typically relies on *perturbation-based diversity* by inducing model disagreement through random feature subsetting or data blinding. While this approximates one notion of epistemic uncertainty, it sacrifices direct characterization of the plausible hypothesis space. We propose the complementary approach: *Rashomon Ensembled Active Learning* (*REAL*) which constructs a committee by exhaustively enumerating the Rashomon Set of all near-optimal models. To address functional redundancy within this set, we adopt a Probably Approximately Correct-Bayesian framework using a Gibbs posterior to weigh committee members by their empirical risk. Leveraging recent algorithmic advances, we exactly enumerate this set for the class of sparse decision trees. Across synthetic and established active learning baselines, REAL outperforms randomized ensembles, particularly in moderately noisy environments where it leverages expanded model multiplicity to achieve faster convergence.

1 INTRODUCTION

Acquiring labeled data is a persistent bottleneck in high-stakes machine learning applications, from image labeling [Huijser and van Gemert, 2017] and sentence classification [Schumann and Rehbein, 2019], to verbal autopsy [McCormick et al., 2016, Fan et al., 2024]. Active Learning (AL) mitigates this cost by strategically querying samples to maximize information gain, thereby achieving high predictive performance with minimal query cost [Settles, 2012, Cohn et al., 1994]. A dominant framework for identifying samples to label is Query-by-Committee (QBC) [Seung et al., 1992, Freund et al., 1997], which selects points where a committee of competing hypotheses exhibits high disagreement.

However, the efficacy of QBC hinges on the models within the committee. Standard approaches, such as Random Forests [Breiman, 2001b], generate committee members via *perturbation-based diversity* by blinding models to specific features, or withholding data through bootstrapping [Hastie et al., 2009]. While bootstrapping offers a principled approximation of one notion of epistemic uncertainty via the sampling variance of the empirical risk minimizer, it sacrifices direct characterization of the statistically plausible hypothesis space to do so. Random feature subsetting compounds this further; while it serves as a powerful decorrelation device to improve ensemble generalization, it acts as a stochastic approximation for diversity rather than a principled mechanism for exploring the hypothesis space. We propose the complementary approach: targeting the uncertainty inherent in model multiplicity by exhaustively enumerating near-optimal models and grounding committee disagreement in their structural ambiguity.

True epistemic uncertainty arises when multiple empirically equivalent hypotheses yield conflicting predictions. This is formalized by the *Rashomon Effect* [Breiman, 2001a], which posits that there exists a multitude of models, the *Rashomon Set*, that explains the data with near-equal accuracy but rely on different structural explanations. Critically, the size of this set has been shown to increase in the presence of label noise [Semenova et al., 2024]. Rather than viewing this expansion as a source of error, we interpret it as an enriched disagreement range that allows our committee to more precisely map the boundaries of the near-optimal hypothesis space.

We hypothesize that an effective active learning strategy should prioritize points that expose structural ambiguity among plausible models, rather than resolving perturbation-based disagreements. To this end, we propose **R**ashomon **E**nsembled **A**ctive **L**earning (**REAL**). Unlike standard ensembles, REAL constructs its committee strictly from the

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4832–4854.

Rashomon set, utilizing a probabilistic weighting scheme to aggregate the competing structural hypotheses into a Rashomon-weighted signal of entropy.

We specifically focus our work on the class of sparse decision trees. This focus is motivated by two factors. First, recent algorithmic advances [Xin et al., 2022, Arslan et al., 2025] allow for the exact and exhaustive enumeration of the Rashomon set for trees. This exact and exhaustive enumeration enables us to capture and incorporate all structurally diverse perspectives into our QBC committee. Second, in the high-stakes domains where active learning is most valuable (e.g., medicine, financial credit reporting, or criminal recidivism prediction), interpretability is non-negotiable; users require models that are not only accurate but also transparent [Semenova et al., 2022, European Parliament and Council, 2024, Pope Leo XIV, 2026]. The interpretability of sparse decision trees is a longstanding pillar of robust statistical learning [Breiman et al., 1984], and modern algorithmic advancements have rigorously formalized their optimization for high-stakes deployment [Lin et al., 2020, Rudin et al., 2022]. By constraining our active learner to this intrinsically interpretable hypothesis class, we guarantee that the final model remains fully transparent to domain experts, bypasses post-hoc explanations, and maintains structural integrity throughout the data-acquisition lifecycle.

Our contributions are as follows: (1) We introduce a querying methodology that replaces perturbation-based diversity with structural diversity, defining the committee strictly as members of the Rashomon set. (2) We integrate a Probably Approximately Correct (PAC)-Bayesian weighting scheme utilizing a Gibbs posterior to resolve functional redundancy within this near-optimal set. (3) We demonstrate that our framework thrives in moderately noisy environments by leveraging expanded pattern diversity. Unlike randomized ensembles that may chase local inconsistencies, our method utilizes the expanded Rashomon set to isolate robust structural patterns, leading to faster and more reliable convergence under label noise. (4) We validate REAL across 20 real-world benchmarks [1]. Our empirical results demonstrate that by grounding query selection in a PAC-Bayesian weighted version space of the Rashomon set, REAL offers a high-precision alternative to stochastic ensembles that is robust to the structural ambiguities of real-world data.

2 RELATED WORK

The core of active learning lies in quantifying uncertainty to distinguish informative signal from uninformative redundancy. Established methods include uncertainty sampling (e.g., Entropy, Least Confidence) [Lewis and Gale, 1994], Expected Model Change (EMC) [Cai et al., 2013], and disagreement-based methods like Query-by-Committee (QBC) [Freund et al., 1997]. In the context of decision trees, QBC is typically implemented using Random Forests or Bagging [Melville and Mooney, 2004, Abe and Mamitsuka, 1998]. While effective, these strategies rely on stochastic randomization that can become unreliable under model misspecification as the committee fails to account for the learner's inherent bias [Burbidge et al., 2007].

To address this vulnerability, we propose a more principled foundation for committee construction. Rather than relying on the randomized bootstrap space, our work constrains the committee to the statistically plausible ϵ-version space defined by the Rashomon set. This shift anchors QBC in the theoretical work of version spaces by Mitchell [1982] as REAL queries prioritize resolving genuine model ambiguity over the variance of random perturbations. While the classical version space of Mitchell [1982] is restricted to hypotheses perfectly consistent with the training data, we define the ϵ-version space as the generalization of this concept to the Rashomon set to encompass all hypotheses within an ϵ-radius of the empirical risk minimizer. This allows our learner to account for model uncertainty even in the presence of noise or misspecification.

The existence of multiple near-optimal models, termed the *Rashomon Effect* by Breiman [2001a], has gained renewed attention in the context of fairness and interpretability. Recent work on predictive multiplicity [Marx et al., 2020, Watson-Daniels et al., 2023] demonstrates that models with identical accuracy can assign vastly different predictions to individual users. While previous studies have studied predictive multiplicity as a source of algorithmic instability [Black et al., 2022, Hsu et al., 2024], REAL utilizes a Gibbs posterior supported by the Rashomon set to transform this multiplicity into a signal for information acquisition. By treating multiplicity as a signal rather than a nuisance, we utilize a PAC-Bayesian weighting scheme to identify regions of epistemic ambiguity where near-optimal hypotheses yield conflicting predictions. Critically, the size of the Rashomon set has been shown to increase in the presence of label noise [Semenova et al., 2024]. Within the REAL framework, we treat this expansion as a source of pattern diversity for our QBC committee [Melville and Mooney, 2004], using the PAC-Bayesian weighting to isolate the most robust structural signals from the resulting larger committee.

Enumerating the full Rashomon set is computationally intensive for many model classes. Pioneering approaches like TreeFarms [Xin et al., 2022] established the methodology for extracting these sets using dynamic programming for sparse decision trees. However, active learning requires recalculating this set iteratively as new labels are acquired. To facilitate this online setting, our work utilizes the recent SORTD algorithm [Arslan et al., 2025], which presents several computational speedups for sparse trees. This efficiency gain allows us to update the exact Rashomon set at every

[1] To ensure reproducibility, our codebase is available at https://github.com/thatswhatsimonsaid/RashomonActiveLearning

step of the active learning loop, scaling the robust analysis of TreeFarms to the iterative active learning setting.

Our weighting strategy is grounded in the framework of PAC-Bayesian learning [McAllester, 1999, Catoni et al., 2008] where we utilize a Gibbs posterior to navigate the structural uncertainty of the hypothesis space. Unlike traditional Bayesian Model Averaging (BMA) [Hoeting et al., 1999] which requires integrating over an often intractable hypothesis space, we adopt a PAC-Bayesian perspective. We utilize the Rashomon set as the empirical support for a Gibbs posterior, capturing the most significant risk-weighted mass of the ϵ-version space without the computational burden of a full Bayesian treatment. By applying a Gibbs distribution to the members of the Rashomon set, REAL implements an approximated Bayesian query strategy.

This approach aligns with the principle of "Occam's Window" of Madigan and Raftery [1994], which suggests that models significantly less plausible than the best-supported model should be discarded to avoid "blurring" the structural signal [Visokay et al., 2025, Venkateswaran et al., 2024]. By constraining our distribution to the Rashomon set, REAL avoids the dilution inherent in full BMA, focusing the learner's attention on the genuine structural contradictions between high-quality hypotheses that represent true epistemic uncertainty. In this regime, the weighted vote entropy reflects the posterior predictive uncertainty within the empirically plausible ϵ-version space, allowing the learner to prioritize queries that resolve genuine structural contradictions rather than stochastic variance.

This focus on structural boundaries aligns tightly with the foundational theory of disagreement-regions and version-space active learning [Hanneke, 2014, Balcan and Feldman, 2013]. By explicitly evaluating the complete Rashomon set, REAL operationalizes these concepts without relying on perturbation-based approximations. Furthermore, our approach contrasts with deep epistemic uncertainty frameworks like BatchBALD [Kirsch et al., 2019], which maximize mutual information over stochastically sampled parameter distributions. Instead, REAL exploits exact structural enumeration to map the version space directly.

3 METHODOLOGY

We consider the standard pool-based active learning setting. Let $\mathcal{X} \in \mathbb{R}^p$ denote the feature space and $\mathcal{Y}$ the label space. We are given a small initial labeled dataset $\mathcal{D}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and a large unlabeled pool $\mathcal{U} = \{\mathbf{x}_j\}_{j=1}^N$. Our objective is to learn a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ from a hypothesis class $\mathcal{H}$ that minimizes generalization risk using a minimal number of queries from $\mathcal{U}$. In this work, $\mathcal{H}$ specifically denotes the class of sparse binary decision trees. While our framework and codebase is designed for multi-class labels, we focus our empirical evaluation on the class of sparse binary decision trees.

In discrete hypothesis spaces like decision trees, the empirical risk minimizer is rarely unique. Instead, there exists a multiplicity of models that fit the training data near-optimally. This phenomenon is formalized as the *Rashomon Effect* [Breiman, 2001a].

Following the formulation in the optimal sparse decision tree literature [Hu et al., 2019, Arslan et al., 2025], we define the objective function for a tree h as a weighted combination of misclassification loss and leaf-count complexity

$$\mathcal{L}(h, \mathcal{D}_{tr}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(h(\mathbf{x}_i) \neq y_i) + \lambda \cdot \text{Leaves}(h) \quad (1)$$

Let h^* denote the optimal tree that minimizes this objective function: $h^* = \text{argmin}_{h \in \mathcal{H}} \mathcal{L}(h, \mathcal{D}_{tr})$.

We define the **ϵ-Rashomon Set**, $\hat{\mathcal{R}}(\epsilon, \mathcal{D}_{tr})$, as the set of all trees with an objective value within a multiplicative factor $1 + \epsilon$ of the optimal model h^* given the training dataset $\mathcal{D}_{tr}$:

$$\hat{\mathcal{R}}(\epsilon, \mathcal{D}_{tr}) = \{h \in \mathcal{H} : \mathcal{L}(h, \mathcal{D}_{tr}) \leq (1 + \epsilon)\mathcal{L}(h^*, \mathcal{D}_{tr})\} \quad (2)$$

Here, $\epsilon \geq 0$ is the *Rashomon Threshold*. When $\epsilon = 0$, the set contains only the exact optimal trees, corresponding to the classical version space described by Mitchell [1982]. As ϵ increases, the set expands to recover the effective ϵ-version space of the problem, retaining diverse models that are indistinguishable based on training error alone. For brevity, we denote the set simply as $\hat{\mathcal{R}}(\epsilon)$ or $\hat{\mathcal{R}}$ when the training data $\mathcal{D}_{tr}$ is clear from context.

3.1 PERTURBED VS. STRUCTURAL DIVERSITY

In the QBC literature, diversity is essential as a committee must represent the full range of disagreement that is actually possible given the current labeled data [Melville and Mooney, 2004]. A committee of identical or highly similar models will offer redundant predictions, failing to identify regions where other equally plausible hypotheses might disagree. The overwhelmingly common approach is to produce this diversity synthetically through resampling procedures [Seung et al., 1992, Abe and Mamitsuka, 1998]. While these methods are theoretically grounded as approximations of the sampling distribution, this approach relies on the stochasticity of the (re)sampling process to induce and explore diverse committee members.

We take a different approach: rather than relying on stochastic perturbations to find diverse models, we enumerate the Rashomon set to exhaustively identify the entire space of hypotheses with high predictive performance. By leveraging the Rashomon effect, where many high-quality models rely on fundamentally different structural explanations for the same data, we provide a committee that represents the inherent model ambiguity of the problem.

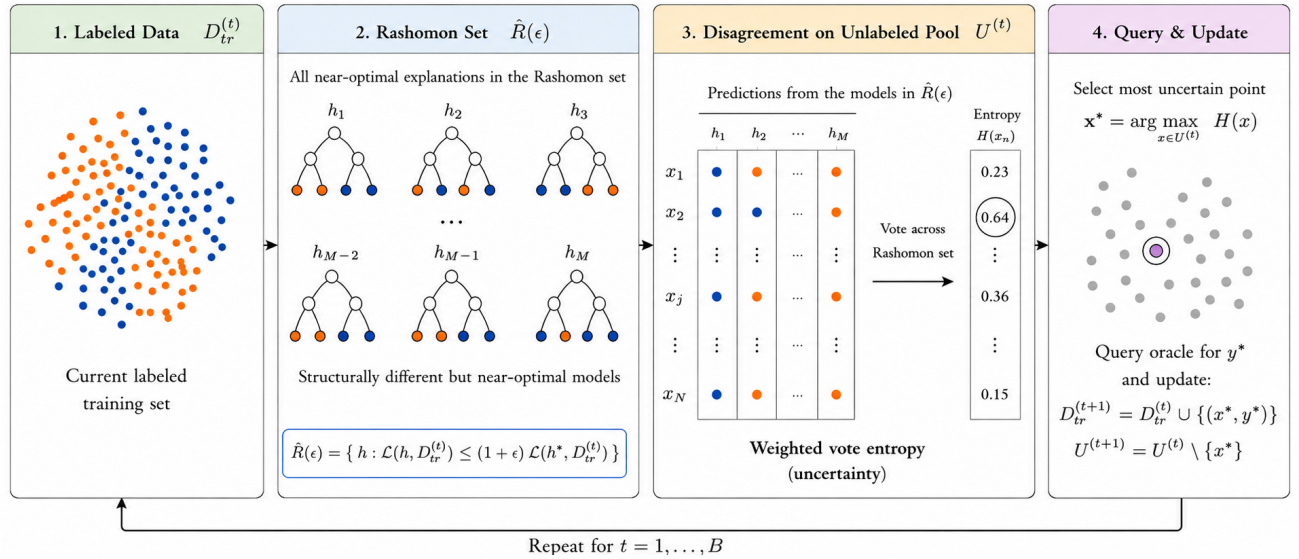


Figure 1: **Rashomon Ensemble Active Learning (REAL).** In each iteration, REAL enumerates the Rashomon set of near-optimal models, computes the weighted vote entropy over the unlabeled pool using Gibbs or uniform weights, and queries the instance maximizing predictive uncertainty before updating the labeled dataset.

3.2 THE REAL ALGORITHM

To ground query selection in genuine model ambiguity, we propose **R**ashomon **E**nsembled **A**ctive **L**earning (**REAL**). Unlike standard ensembles that rely on a fixed number of randomized members, REAL utilizes the membership of the ϵ-Rashomon set $\hat{\mathcal{R}}(\epsilon, \mathcal{D}_{tr})$ to characterize the ϵ-version space. By requiring that every committee member satisfies the Rashomon bound, we ensure that any observed disagreement represents a conflict between two plausible structural hypotheses grounded in the same evidence.

While the Rashomon set provides a comprehensive view of the ϵ-version space, naively treating every member with equal weight can lead to structural saturation, where minor structural variants of the same hypothesis dominate the ensemble vote. This leads to a loss of effective diversity, which is critical for the success of QBC as a diverse ensemble is necessary to accurately capture the boundaries of the ϵ-version space [Melville and Mooney, 2004].

Although there are many ways to prune a set of trees to a unique set of representative structures, such as by agglomerative clustering or predictive equivalence [McTavish et al., 2025], pruning ignores the relative empirical support of different hypotheses. However, naively treating the raw density of the ϵ-version space as a signal of evidence is equally problematic. In tree-based models, density is often distorted by structural artifacts: more complex trees naturally exhibit higher multiplicity simply due to the combinatorial permutations of their splits. Rather than discarding models via pruning or allowing this structural bloat to artificially dominate the ensemble vote, our PAC-Bayesian weighting scheme corrects for this imbalance. Because our Gibbs posterior, formally introduced below, is defined over the regularized objective $\mathcal{L}(h, \mathcal{D}_{tr})$, the complexity penalty aggressively down-weights these structurally complex, redundant variants. This ensures that the query strategy is informed by genuine empirical risk within the Rashomon set.

We define a Gibbs posterior supported on the membership of $\hat{\mathcal{R}}(\epsilon, \mathcal{D}_{tr})$. Following the literature on PAC-Bayesian learning [Catoni et al., 2008], we assign each tree $h \in \hat{\mathcal{R}}(\epsilon)$ a Gibbs weight w_h based on its empirical risk:

$$w_h = \frac{\exp(-\beta \cdot \mathcal{L}(h, \mathcal{D}_{tr}))}{\sum_{h' \in \hat{\mathcal{R}}} \exp(-\beta \cdot \mathcal{L}(h', \mathcal{D}_{tr}))} \tag{3}$$

where $\beta \geq 0$ is an inverse temperature parameter that controls the "peakedness" of the distribution. With this weighting, trees that fit the training evidence more precisely have a greater influence on the query decision, while the threshold ϵ ensures that the ensemble maintains sufficient diversity to explore the ϵ-version space.

Based on the choice of weight distribution, we define two specific implementations of the framework: (i) **UNREAL** (Uniform REAL), a baseline utilizing a maximum-entropy prior where $w_h = 1/|\hat{\mathcal{R}}|$ for all $h \in \hat{\mathcal{R}}$, and (ii) **BREAL** (Bayesian REAL), our primary methodology which leverages the Gibbs posterior to resolve functional redundancy and prioritize models with higher empirical support.

Given the weighted ensemble, we select the query instance $\mathbf{x}^*$ that maximizes the *Weighted Vote Entropy*:

$$P(y \mid \mathbf{x}, \hat{\mathcal{R}}) = \sum_{h \in \hat{\mathcal{R}}} w_h \cdot \mathbb{I}(h(\mathbf{x}) = y) \tag{4}$$

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathcal{U}}{\operatorname{argmax}} \left[-\sum_{y\in\mathcal{Y}} P(y \mid \mathbf{x}, \hat{\mathcal{R}}) \log P(y \mid \mathbf{x}, \hat{\mathcal{R}}) \right]. \quad (5)$$

This formulation allows for a continuous measure of disagreement that respects the likelihood of each hypothesis.

To quantify the richness of the ϵ-version space during active learning, we calculate the Effective Committee Size (ECS):

$$ECS(\hat{\mathcal{R}}) = \exp\left\{ -\sum_{h\in\hat{\mathcal{R}}} w_h \ln w_h \right\} \quad (6)$$

The ECS provides a diagnostic of model certainty: a high ECS indicates a diverse ϵ-version space where many theories are equally plausible, while an ECS approaching one signals a collapse of the ϵ-version space onto a single hypothesis.

REAL targets points where the set of all near-optimal models exhibits maximal disagreement. As new labels y^* are acquired, the Rashomon set is updated and the additional information prunes hypotheses inconsistent with the observed data. This process systematically narrows the ϵ-version space, concentrating the learner's attention on the structural features of the hypothesis manifold that remain empirically plausible. For implementation details, the complete REAL procedure is formally defined in Algorithm 1 (Appendix A).

4 THE BOUNDARIES OF PRECISION: MISSPECIFICATION AND NOISE

To understand the robustness and limitations of Rashomon-based active learning, we systematically explore the impact of model misspecification and label noise.

4.1 EXPERIMENTAL DESIGN

We generate synthetic datasets $\mathcal{D} = \{\mathbf{X}, \mathbf{Y}\}$ where $\mathbf{X} \in \{0,1\}^d$ is sampled from a Bernoulli(0.5) distribution. The labeling process $\mathbf{Y}$ follows a mixture model defined by $\alpha \in [0,1]$, which interpolates between a non-linear XOR pattern (dependent on the first two features) and a linear threshold model (dependent on all features weighted by $\mathbf{w} \sim \mathcal{N}(0,1)$). To evaluate the impact of label uncertainty, we introduce flip-noise $\phi \in [0, 0.5]$ to the target vector, where each label is inverted with probability ϕ.

We control model misspecification via mixture parameter α for a Bernoulli formulation:

$$Y_i = S_i \cdot y_{\text{linear},i} + (1 - S_i) \cdot y_{\text{tree},i}, \quad S_i \sim \text{Bernoulli}(\alpha) \quad (7)$$

This setup creates a deliberate "needle-in-a-haystack" problem specifically designed to challenge the stochastic construction of standard ensembles. In a high-dimensional space where only two features are informative and the remaining 18 act as independent noise, the random feature subsetting and bootstrapping inherent to QBC-RF often produce committee members that miss the structural signal entirely by chance. In contrast, REAL utilizes an exhaustive search of the ϵ-version space to identify the sparse set of models that remain consistent with the empirical evidence, allowing it to isolate the informative features more efficiently. As $\alpha \to 1$, however, the committee members of both REAL and QBC-RF become fundamentally misspecified as individual estimators struggle to approximate the diagonal linear boundary with axis-aligned splits.

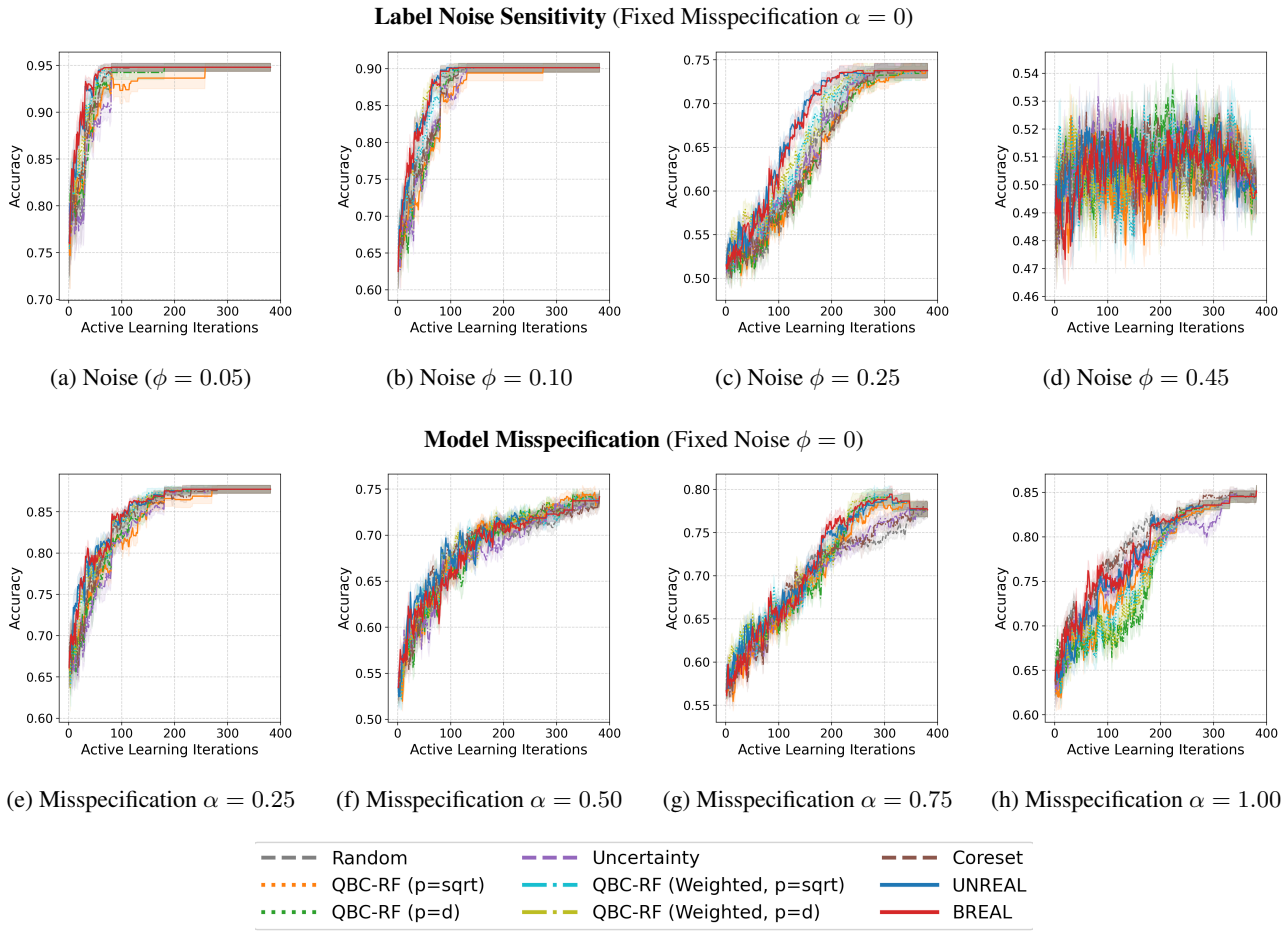
In a separate experimental regime, we investigate the impact of stochastic corruption by applying symmetric label noise, flipping a fraction $\phi \in \{0.05, 0.10, 0.25, 0.45\}$ of labels. Our motivation for this study is grounded in the work of Semenova et al. [2024], who demonstrate that increased noise fundamentally expands the Rashomon set and increases *pattern diversity*, the average difference in predictions among near-optimal models. This experiment tests whether the active learner can distinguish the underlying structural signal from the combinatorial explosion of noise-induced hypotheses that emerge as the ϵ-version space expands.

4.2 EXPERIMENTAL PROTOCOL

Experiments used $N = 25$ independent replications with randomized initial seeds. For each run, we held out 20% of the observations as a static test set to evaluate predictive performance at each iteration. The remaining 80% constitutes the working pool from which the learner begins with a stratified initial set of 20 samples.

To distinguish prediction from acquisition, we define two model roles: the *Predictor* (maximum depth 5, $\lambda = 0.001$) is used for final test-set evaluation, while the *Selector* characterizes the ϵ-version space for query selection. We tune the *Selector* over a restricted depth to prioritize robust structural patterns over aleatoric noise. By restricting depth ($d \in \{3, 5\}$) and increasing complexity penalties ($\lambda \in \{0.01, 0.001\}$), we create a structural bottleneck that forces the learner to resolve global topological ambiguities rather than exhausting the budget on local inconsistencies that deeper, unregularized models might overfit.

We calibrate the *Selector* on the initial 20-sample pilot set $\mathcal{D}_{tr}^{(0)}$ to ground its parameters: (i) depth $d \in \{3, 5\}$ and complexity $\lambda \in \{0.01, 0.001\}$ are selected via leave-one-out cross-validation; (ii) the Rashomon radius ϵ is iteratively expanded in 0.05 increments until $|\hat{\mathcal{R}}| \geq 2$ (the floor for committee-based disagreement [Seung et al., 1992]) and (iii) β is optimized via grid search over $\{1.0, \ldots, 500.0\}$ to maximize the Area Under the Precision-Recall Curve (AUPRC) between vote entropy and empirical classification errors. This ensures the uncertainty signal is a calibrated proxy for epistemic ambiguity. These optimal values are applied to all *Selector* models to ensure a fair comparison.



(a) Noise ($\phi = 0.05$) (b) Noise $\phi = 0.10$ (c) Noise $\phi = 0.25$ (d) Noise $\phi = 0.45$

(e) Misspecification $\alpha = 0.25$ (f) Misspecification $\alpha = 0.50$ (g) Misspecification $\alpha = 0.75$ (h) Misspecification $\alpha = 1.00$

Figure 2: (Top) REAL leverages noise-induced version-space diversity to maintain superior performance, with its advantage widening until extreme noise floors ($\phi = 0.45$) equalize all strategies. (Bottom) Under total misspecification ($\alpha = 1.0$), all tree-based QBC strategies underperform relative to Passive Sampling. However, BREAL and UNREAL demonstrate unique resilience, outperforming the other QBC methods and maintaining early-stage efficiency despite the structural misalignment.

Appendix D provides a threshold sensitivity ablation study of BREAL's robustness to the choice of ϵ.

Comparison methods and hyperparameters are selected to isolate version-space precision from perturbation-based diversity. We include two variants of QBC-RF utilizing feature subsets of size $m = \sqrt{d}$ and $m = d$ both with 100 base learners. To control for the impact of weighting, we also apply both Uniform and PAC-Bayesian weighting to QBC-RF. This isolates whether the benefits of the Gibbs posterior are universal across ensemble types or specifically synergistic with the structural geometry of the Rashomon set.

Within the REAL framework, we evaluate two distinct configurations to isolate the influence of version-space weighting from the underlying topology. Our primary methodology, BREAL, applies the Gibbs posterior (Equation 3) to the committee, weighting each member's vote by its empirical risk to resolve functional redundancy within the ϵ-version space. In contrast, UNREAL serves as a structural baseline that assumes a uniform distribution over the hypothesis space, treating all near-optimal hypotheses as equally informative. By assigning equal weight to all near-optimal hypotheses, it allows us to quantify the performance gains specifically attributable to the PAC-Bayesian weighting scheme.

We benchmark REAL against Uncertainty Sampling [Settles, 2012], Coreset Diversity [Sener and Savarese, 2018], and QBC-RF variants ($m \in \{\sqrt{d}, d\}$), each with Uniform and PAC-Bayesian weighting. Uncertainty Sampling uses least-confidence scoring on an optimal tree. For Coreset, we implement the greedy k-center algorithm using Euclidean distance in the feature space to ensure geometric coverage. These baselines ensure that the REAL provides a meaningful improvement over industry standards. An overview of methods is provided in Table 1 of Appendix A.

4.3 SYNTHETIC RESULTS

Misspecification Gap. Our results reveal a nuanced relationship between version-space precision and model misspecifi-

cation. In the settings of low misspecification ($\alpha = 0.25$), REAL (both versions) demonstrates superiority (see Figures 2 and 3), reaching a higher accuracy ceiling while standard ensembles struggle to identify the sparse structural signal.

As we introduce misspecification ($\alpha = \{0.50, 0.75, 1.00\}$), the performance gap narrows. Yet, REAL demonstrates unique resilience. In these regimes, the Rashomon set effectively becomes a collection of "equally misspecified experts." While REAL's committee becomes hypersensitive to local misalignments, treating every structural disagreement as a reason to query a new point, this exhaustive search ensures that the learner identifies the most plausible (albeit biased) solution faster than QBC-RF baselines. In contrast, the algorithmic variance of QBC-RF acts as a structural buffer. By subsetting features and bootstrapping data, individual trees become diverse across different subsets of noise, allowing the ensemble average to effectively smooth over the geometric confusion of the misspecification.

The advantage of this structural search is most apparent under total misspecification ($\alpha = 1.0$). While the smoothing effect of QBC-RF eventually stabilizes, it requires significantly more samples to resolve. In contrast, REAL consistently outperforms QBC-RF baselines by identifying stable, near-optimal decision boundaries much earlier in the learning process. This indicates that even when the hypothesis space cannot perfectly represent the target function, an exhaustive search of the Rashomon set is a more efficient path to the accuracy ceiling than relying on the stochastic variance of feature subsetting or bootstrapping.

Leveraging Noise. REAL's performance advantage widens with label noise (ϕ). While standard randomized ensembles exhibit significant instability at $\phi = 0.25$, REAL reaches its accuracy plateau nearly 100 iterations earlier.

We attribute this to the fundamental relationship between noise and version-space geometry. As noted by Semenova et al. [2024], label noise expands the Rashomon set, thereby increasing the pattern diversity of near-optimal models. REAL strategically leverages this expanded Rashomon set to construct a more diverse committee that captures a wider array of structural explanations. The Gibbs-weighted framework of BREAL then acts as a probabilistic filter for the expanding set of explanations, allowing the learner to explore the diversity of the expanding hypotheses space while simultaneously de-prioritizing those that overfit outliers.

In this setting, moderate noise works in REAL's favor by providing a broader range of perspectives for the committee. At extreme noise ($\phi = 0.45$), however, the advantage evaporates as the structural signal becomes indistinguishable from aleatoric chaos. These results suggest that while stochastic ensembles treat noise as a source of instability, REAL transforms it into a signal of structural diversity that identifies the most robust patterns in the ϵ-version space.

5 BENCHMARK EXPERIMENTS

To evaluate the practical utility of REAL, we conduct an extensive study across 20 real-world datasets (Table 2). We maintain the exact experimental protocol as in Section 4.2.

To provide a global assessment of performance across our 20 benchmarks, we utilize the Efficiency Ratio based on the Area Under the Curve (AUC). Since active learning methods converge as budgets exhaust, total AUC dilutes the signal of early-stage efficiency. We therefore report AUC truncated at 70% to focus on the window of maximal information discovery where labeling constraints are most acute.

To quantify cumulative performance, we calculate the Truncated AUC using the trapezoidal rule. For a budget fraction $K \in (0, 1]$, we define the truncation point $T = \lfloor K \cdot N \rfloor$. The AUC for method m is given by:

$$\text{AUC}_{m,K} = \sum_{t=1}^{T-1} \frac{A_m(t) + A_m(t+1)}{2} \tag{8}$$

where $A_m(t)$ denotes the mean accuracy of method m at iteration t. We then define the Efficiency Ratio (ρ) relative to the BREAL baseline:

$$\rho_m = \frac{\text{AUC}_{m,K}}{\text{AUC}_{\text{BREAL},K}} \tag{9}$$

In this metric, a ratio $\rho < 1.0$ indicates that BREAL captures a greater area under the learning curve, demonstrating higher accuracy during the initial 70% of the labeling budget. Our AUC evaluation protocol specifically computes the area under the standard metric convergence curve, following the evaluation framework established in recent large-scale active learning benchmarking suites [Bi et al., 2025]

5.1 BENCHMARK EXPERIMENT RESULTS

Accuracy. Figure 3 presents the Efficiency Ratio of each method relative to BREAL (values < 1.0 in green indicate that BREAL achieved superior sample efficiency). Results demonstrate that BREAL consistently outperforms Random Sampling and Uncertainty Sampling across a diverse range of geometries. This advantage is particularly pronounced in high-multiplicity benchmarks like *Tic-Tac-Toe* and *Hepatitis* (see Table 2), where REAL's exhaustive version-space characterization prevents the learner from being distracted by the algorithmic noise that hampers randomized ensembles.

While the truncated AUC provides a global summary of efficiency, the specific learning dynamics of the ϵ-version space, including the rate of accuracy growth and the stabilization of the model structure, are best observed through individual learning curves. We provide the complete set of Accuracy along with other metrics (e.g. Tree Edit Distance) trace plots for all 20 benchmarks Appendix B.

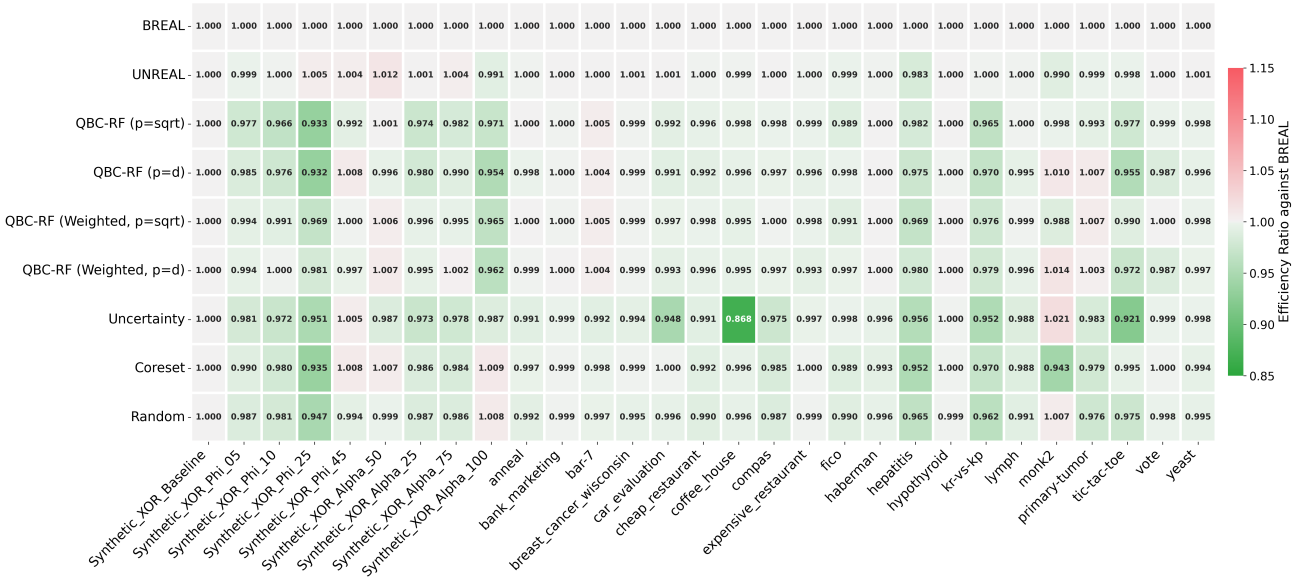


Figure 3: **Active Learning Efficiency Heatmap** ($K = 0.7$). This figure displays the Efficiency Ratio (ρ) of standard baselines relative to BREAL, with values below 1.0 (**Green**) indicating superior sample efficiency for our method. The results demonstrate that REAL reaches peak predictive performance significantly faster than randomized ensembles.

We track the rate of version-space collapse via the ECS (Equation 6). As evidenced in the trace plots of Appendix B, this metric visualizes how quickly the structural signal is resolved across different noise regimes.

Label Efficiency. Practical deployment of active learning strategies often hinges on cost reduction. We quantify this using *Relative Label Efficiency* (N_{rel}): the ratio of labels required by a method to reach a specific performance milestone relative to a Random Sampling baseline. We define these milestones as 70%, 80%, and 90% of the total accuracy growth achieved over the labeling budget. An $N_{rel} < 1.0$ indicates superior efficiency. Figure 4 aggregates these scores across the benchmark suite. Our results show BREAL frequently achieves the 70% milestone with less than half the labels required by passive sampling.

The results reveal two critical advantages of REAL. Firstly, both REAL variants consistently shift the median efficiency distribution below 0.5, implying a reduction in labeling costs of over 50% on many datasets as opposed to Random Sampling. For BREAL, the median remains exceptionally stable across milestones, even as the labeling budget tightens for the 90% growth target.

Secondly, a major finding is the extreme volatility of advanced baselines that is not present in REAL. While QBC-RF methods and Uncertainty Sampling achieve competitive efficiency on select datasets (left whiskers extending toward 0.1), they exhibit heavy tails of extreme inefficiency, with right whiskers and outliers extending beyond 2.0. This indicates that these methods can often be twice as expensive as random sampling when their stochastic or greedy heuristics fail to capture the data geometry. In contrast, REAL maintains a tight, compact distribution with minimal right skew and no extreme outliers. This high-precision, low-variance performance makes REAL a safer and more predictable choice for real-world applications where labeling budgets are fixed and dataset characteristics are initially unknown.

The PAC-Bayesian Advantage. Our results reveal an improvement over standard QBC when incorporating PAC-Bayesian weighting. As shown in Figures 3, 4 and the trace plots of Appendix B, applying a Gibbs posterior to standard ensembles (QBC-RF Weighted) already improves efficiency over their uniform counterparts (QBC-RF Uniform).

However, the most significant performance gains are realized only when moving from randomized ensembles to the version-space architecture of the REAL framework. While a Gibbs posterior weighting refines randomized committees, the Rashomon set's structural diversity provides a more direct characterization of version-space uncertainty than synthetic RF diversity. BREAL consistently establishes a higher accuracy ceiling several iterations earlier than even the QBC-RF weighted baselines as evidenced in the as *Anneal*, *Hepatitis*, and *Tic-Tac-Toe* datasets with accuracy gains of up to 0.045 over conventional QBC-RF and 0.031 over its PAC-Bayesian weighted counterpart.

Crucially, this weighting also acts as a robust structural filter against threshold misspecification. As detailed in the ablation study in Appendix D, BREAL's Gibbs posterior ensures the active learning signal remains securely anchored to the high-evidence core of the version space even when the ϵ-boundary is aggressively relaxed.

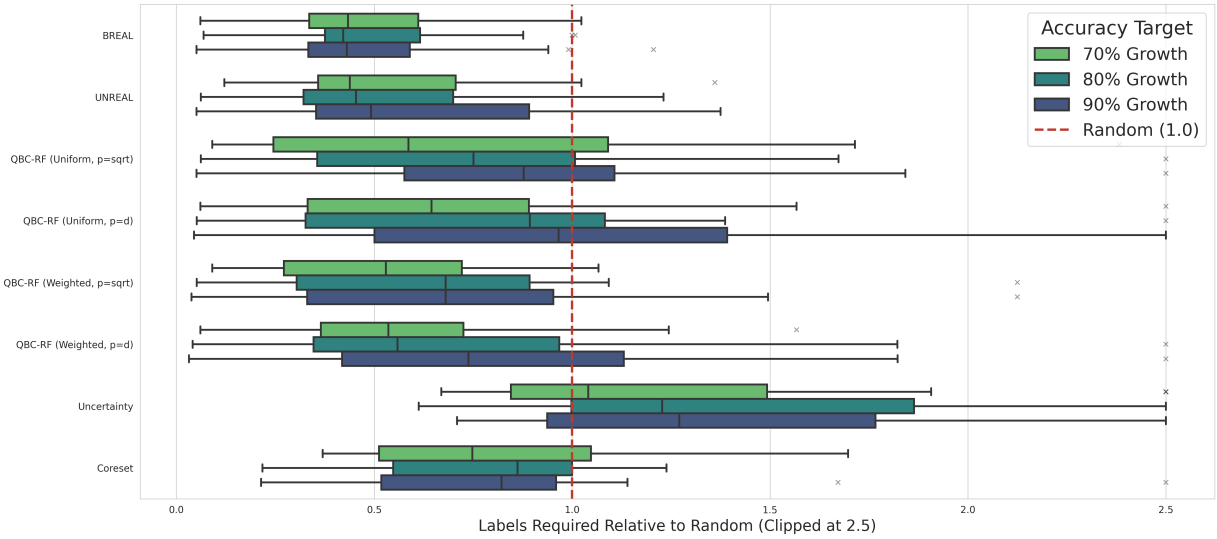


Figure 4: **Relative Label Efficiency (N_{rel}) across 20 datasets.** Each boxplot summarizes the labeling budget required to achieve 70%, 80%, and 90% of the total possible accuracy gain. Values to the left of the dashed line (< 1.0) indicate superior efficiency relative to Random Sampling. Both UNREAL and BREAL maintain a significantly tighter distribution and lower median than standard baselines, demonstrating that characterizing the Rashomon set provides a more reliable and robust path to cost reduction than randomized ensembles or greedy heuristics.

Computational Complexity

Table 3 in the appendix presents the median runtime across 25 independent simulations for each dataset. As expected, the REAL variants incur a higher computational cost compared to randomized ensembles. This overhead is a direct consequence of the exhaustive version-space characterization performed by the SORTD algorithm at each iteration.

While this enumeration carries a computational premium over perturbation-based alternatives, the limitation is a practical trade-off given the setting of active learning. In high-stakes applications (e.g., medical diagnostics or financial credit evaluation), the cost of acquiring a single expert label often overshadows CPU runtimes. In these regimes, sample efficiency is a key constraint, making the tradeoff of runtime for reduced labeling costs a reasonable consideration.

6 CONCLUSION

This work introduced REAL, an active learning framework that replaces the perturbation-based diversity of randomized ensembles with the principled diversity of the Rashomon set. By constructing committees from near-optimal models, REAL targets true epistemic uncertainty, allowing for superior sample efficiency in well-specified regimes.

Our empirical analysis disentangles the roles of version-space precision and algorithmic noise. We demonstrate that rather than being a vulnerability, moderate label noise drives pattern diversity strategically leveraged by REAL. By utilizing a Gibbs-weighted framework, the learner navigates the expanded ϵ-version space to find a stable structural signal. While the perturbation-based diversity of Random Forests provides a necessary structural buffer under geometric misspecification, REAL remains a superior instrument for high-precision structural discovery in well-aligned regimes.

Several directions remain open. While our implementation focuses on sparse decision trees, the most immediate frontier is the extension of REAL to other hypothesis classes. The ongoing expansion of Rashomon enumeration to structures such as Rashomon Partitions [Venkateswaran et al., 2024] and Prototypical-Part Networks [Donnelly et al., 2025] suggests that REAL can be adapted to more complex model classes. Additionally, for extreme architectural scales or massive datasets where computation remains a severe bottleneck for exact enumeration, future work will focus on integrating our modular framework with recent anytime approximation schemes, such as PRAXIS [Heile et al., 2026]. This extension would trade strict mathematical exhaustiveness for immediate computational scalability, broadening the applicability of REAL while addressing the heavy computational complexity.

By integrating near-optimal models rather than searching for the "best model," we provide a more grounded path toward interpretable and efficient machine learning that directly addresses the challenge of quantifying epistemic uncertainty in the presence of unresolvable model ambiguity.

References

Naoki Abe and Hiroshi Mamitsuka. Query learning strategies using boosting and bagging. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 1–9, San Francisco, CA, USA, 1998.

Elif Arslan, Jacobus G. M. van der Linden, Serge Hoogendoorn, Marco Rinaldi, and Emir Demirović. Sorted rashomon sets of sparse decision trees: Anytime enumeration. In *Advances in Neural Information Processing Systems (NeurIPS 2025) Poster Session*, 2025. San Diego, CA.

Maria-Florina F Balcan and Vitaly Feldman. Statistical active learning algorithms. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, 2013.

Jinghou Bi, Yuanhao Xu, Felix Conrad, Hajo Wiemer, and Steffen Ihlenfeldt. A comprehensive benchmark of active learning strategies with automl for small-sample regression in materials science. *Scientific Reports*, 15, 2025.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 850–863, New York, NY, USA, 2022. doi: 10.1145/3531146.3533149.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984.

Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199 – 231, 2001a. doi: 10.1214/ss/1009213726.

Leo Breiman. Random forests. *Machine Learning*, 45(1): 5–32, 2001b. doi: 10.1023/A:1010933404324.

Robert Burbidge, Jem J. Rowland, and Ross D. King. Active learning for regression based on query by committee. In *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning*, IDEAL'07, page 209–218, Berlin, Heidelberg, 2007.

Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, 2013. doi: 10.1109/ICDM.2013.104.

O. Catoni, Project Euclid, Cornell University. Library, and Duke University Press. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Lecture notes-monograph series. Cornell University Library, 2008.

David Cohn, Zoubin Ghahramani, and Michael Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, 1994.

Jon Donnelly, Zhicheng Guo, Alina Jade Barnett, Hayden McTavish, Chaofan Chen, and Cynthia Rudin. Rashomon sets for prototypical-part networks: Editing interpretable models in real-time. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4528–4538, 2025. doi: 10.1109/CVPR52734.2025.00427.

European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council (artificial intelligence act). Official Journal of the European Union, L 2024/1689, 2024.

Shuxian Fan, Adam Visokay, Kentaro Hoffman, Stephen Salerno, Li Liu, Jeffrey T. Leek, and Tyler H. McCormick. From narratives to numbers: Valid inference using language model predictions from verbal autopsy narratives. *CoRR*, abs/2404.02438, 2024. doi: 10.48550/ARXIV.2404.02438.

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Mach. Learn.*, 28:133–168, September 1997. ISSN 0885-6125. doi: 10.1023/A:1007330508534.

Steve Hanneke. Theory of disagreement-based active learning. *Found. Trends Mach. Learn.*, 7(2–3):131–309, June 2014. ISSN 1935-8237. doi: 10.1561/2200000037.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer New York Inc., 2 edition, 2009.

Zakk Heile, Hayden McTavish, Varun Babbar, Margo Seltzer, and Cynthia Rudin. From rashomon theory to praxis: Efficient decision tree rashomon sets, 2026.

Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, 14(4):382 – 417, 1999. doi: 10.1214/ss/1009212519.

Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Chun-Fu Chen. Rashomongb: analyzing the rashomon effect and mitigating predictive multiplicity in gradient boosting. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024.

Xiyang Hu, Cynthia Rudin, and Margo Seltzer. Optimal sparse decision trees. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Miriam Huijser and Jan C. van Gemert. Active decision boundary annotation with deep generative models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In Bruce W. Croft and C. J. van Rijsbergen, editors, *SIGIR '94*, pages 3–12, London, 1994.

Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and scalable optimal sparse decision trees. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020.

David Madigan and Adrian E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994. doi: 10.1080/01621459.1994.10476894.

Charles T. Marx, Flavio Du Pin Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020.

David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, page 164–170, New York, NY, USA, 1999. Association for Computing Machinery. doi: 10.1145/307400.307435.

Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

Hayden McTavish, Zachery Boner, Jon Donnelly, Margo Seltzer, and Cynthia Rudin. Leveraging predictive equivalence in decision trees, 2025.

Prem Melville and Raymond J. Mooney. Diverse ensembles for active learning. In *Proceedings of 21st International Conference on Machine Learning (ICML-2004)*, pages 584–591, Banff, Canada, July 2004.

Tom M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(82)90040-6.

Pope Leo XIV. *Magnifica Humanitas: Encyclical on Safeguarding the Human Person in the Time of Artificial Intelligence*. Melville House, 2026. ISBN 9781685893095.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 01 2022. doi: 10.1214/21-SS133.

Raphael Schumann and Ines Rehbein. Active learning via membership query synthesis for semi-supervised sentence classification. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 472–481, Hong Kong, China, November 2019. doi: 10.18653/v1/K19-1044.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1827–1858, New York, NY, USA, 2022. doi: 10.1145/3531146.3533232.

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Burr Settles. *Active Learning*. Morgan & Claypool Publishers, 2012.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA, 1992. doi: 10.1145/130385.130417.

Aparajithan Venkateswaran, Anirudh Sankar, Arun G. Chandrasekhar, and Tyler H. McCormick. Robustly estimating heterogeneity in factorial data using rashomon partitions. *ArXiv*, abs/2404.02141, 2024.

Adam Visokay, Eddie Hock, Kentaro Hoffman, and Tyler H McCormick. When you come to a fork in the road, take it: the rashomon effect for social scientists, Apr 2025.

Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the Thirty-Seventh AAAI Conference on*

Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23, 2023. doi: 10.1609/aaai.v37i9.26227.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022.

Supplementary Material

Simon D. Nguyen[1] **Hayden McTavish**[3] **Kentaro Hoffman**[1] **Tyler H. McCormick**[1,2] **Cynthia Rudin**[3]

[1]Department of Statistics, University of Washington
[2]Department of Sociology, University of Washington
[3]Department of Computer Science, Duke University

A METHODS AND IMPLEMENTATION DETAILS

Algorithm 1 Rashomon Ensembled Active Learning (REAL)

Input: Initial set $\mathcal{D}_{tr}^{(0)}$; Pool $\mathcal{U}^{(0)}$; Budget B;
    Rashomon threshold ϵ; Weighting mode $\omega \in \{\text{Gibbs, Uniform}\}$
Initialize: Selector parameters (d, λ, β) via multi-stage calibration on $\mathcal{D}_{tr}^{(0)}$
for $t = 0$ **to** B **do**
    1. Version-Space Characterization
    $h^* \leftarrow \text{OptimalTree}(\mathcal{D}_{tr}^{(t)}, d, \lambda)$
    $\hat{\mathcal{R}}(\epsilon) \leftarrow \text{SORTD_Enumerate}(h^*, \epsilon, \mathcal{D}_{tr}^{(t)})$

    2. PAC-Bayesian Weighting
    for each $h_i \in \hat{\mathcal{R}}(\epsilon)$ **do**
        if $\omega = \text{Gibbs}$ **then**
            $w_i \leftarrow \frac{\exp(-\beta \cdot \mathcal{L}(h_i, \mathcal{D}_{tr}^{(t)}))}{\sum_j \exp(-\beta \cdot \mathcal{L}(h_j, \mathcal{D}_{tr}^{(t)}))}$ {**B-REAL** configuration}
        else
            $w_i \leftarrow 1/|\hat{\mathcal{R}}(\epsilon)|$ {**UNREAL** configuration}
        end if
    end for
    $\text{ECS}^{(t)} \leftarrow \exp(-\sum w_i \ln w_i)$

    3. Epistemic Uncertainty Quantification
    for each candidate $\mathbf{x}_n \in \mathcal{U}^{(t)}$ **do**
        $\hat{P}(y|\mathbf{x}_n) \leftarrow \sum_{h_i \in \hat{\mathcal{R}}} w_i \cdot \mathbb{I}(h_i(\mathbf{x}_n) = y)$
        $H_n \leftarrow -\sum_{y \in \mathcal{Y}} \hat{P}(y|\mathbf{x}_n) \log \hat{P}(y|\mathbf{x}_n)$
    end for

    4. Label Acquisition and Update
    $\mathbf{x}^* \leftarrow \arg\max_{\mathbf{x}_n \in \mathcal{U}^{(t)}} H_n$
    $y^* \leftarrow \text{QueryOracle}(\mathbf{x}^*)$
    $\mathcal{D}_{tr}^{(t+1)} \leftarrow \mathcal{D}_{tr}^{(t)} \cup \{(\mathbf{x}^*, y^*)\}; \mathcal{U}^{(t+1)} \leftarrow \mathcal{U}^{(t)} \setminus \{\mathbf{x}^*\}$
end for
$f \leftarrow \text{OptimalTree}(\mathcal{D}_{tr}^{(B)}, d_{\text{pred}}, \lambda_{\text{pred}})$
Return: Predictor f,

Label	Committee Foundation	Weighting	Diversity Mechanism
Random	Random Forest	Uniform	Passive (Random Baseline)
QBC-RF ($m = \sqrt{d}$)	Random Forest	Uniform	Artificial (Feature Blinding)
B-QBC-RF ($m = \sqrt{d}$)	Random Forest	Gibbs	Artificial + PAC-Bayesian
QBC-RF ($m = d$)	Random Forest	Uniform	Artificial (Bootstrap Only)
B-QBC-RF ($m = d$)	Random Forest	Gibbs	Artificial + PAC-Bayesian
Uncertainty	Optimal Sparse Tree	—	Greedy
Coreset	Optimal Sparse Tree	—	Geometric (k-Center)
UNREAL	Rashomon Set	Uniform	Version-Space
BREAL	Rashomon Set	Gibbs	Version-Space

Table 1: Overview of evaluated active learning configurations. We categorize strategies by their committee foundation and weighting scheme to isolate the performance gains of structural version-space characterization from those of PAC-Bayesian posterior weighting.

B BENCHMARK DETAILS

No.	Dataset	Src	N	d	$\lvert\hat{\mathcal{R}}\rvert$	ECS
1	Anneal	UCI	812	44	2	2.0
2	Bank Marketing	UCI	4,521	23	505	505.0
4	Bar-7	SORTeD	1,913	14	24	24.0
5	Breast Cancer WI	UCI	699	10	67	67.0
6	Car Evaluation	UCI	1,728	15	264	263.6
7	Cheap Restaurant	SORTeD	2,653	15	4	4.0
8	Coffee House	SORTeD	3,816	15	1,633	1,631.6
9	COMPAS	ProPublica	6,907	12	30,943	30,850.4
10	Expensive Restaurant	SORTeD	1,417	15	3,869	3,868.8
11	FICO (HELOC)	FICO	10,459	17	100,106	100,049.1
12	Haberman	UCI	306	92	7	7.0
13	Hepatitis	UCI	137	34	9,469	9,463.1
14	Hypothyroid	UCI	3,247	39	1	1.0
15	Lymphography	UCI	148	47	104	104.0
16	MONK-2	UCI	169	11	1,229	1,225.1
17	Primary Tumor	UCI	336	17	4,022	4,019.9
18	Tic-Tac-Toe	UCI	958	18	311	310.7
19	Congressional Vote	UCI	435	48	2	2.0
20	Yeast	UCI	1,484	46	1,046	1,045.9
21	Synth. XOR (α=0, ϕ=0)	Syn	500	20	2	2.0
22	Synth. XOR (α=0.25)	Syn	500	20	794	794.0
23	Synth. XOR (α=0.50)	Syn	500	20	2,292	2,290.0
24	Synth. XOR (α=0.75)	Syn	500	20	64	63.8
25	Synth. XOR (α=1.00)	Syn	500	20	12	12.0
26	Synth. XOR (ϕ=0.05)	Syn	500	20	794	794.0
27	Synth. XOR (ϕ=0.10)	Syn	500	20	794	794.0
28	Synth. XOR (ϕ=0.25)	Syn	500	20	989	986.2
29	Synth. XOR (ϕ=0.45)	Syn	500	20	100,001	99,897.3

Table 2: Summary of benchmark datasets and version space characteristics. N and d denote the number of samples and features, respectively. $|\hat{\mathcal{R}}|$ represents the size of the Rashomon set (the number of models within ϵ of the empirical risk minimizer), and ECS denotes the Effective Committee Size, quantifying the diversity of plausible theories within the Rashomon set as defined by the Gibbs distribution. All metrics are computed globally across the entire dataset.

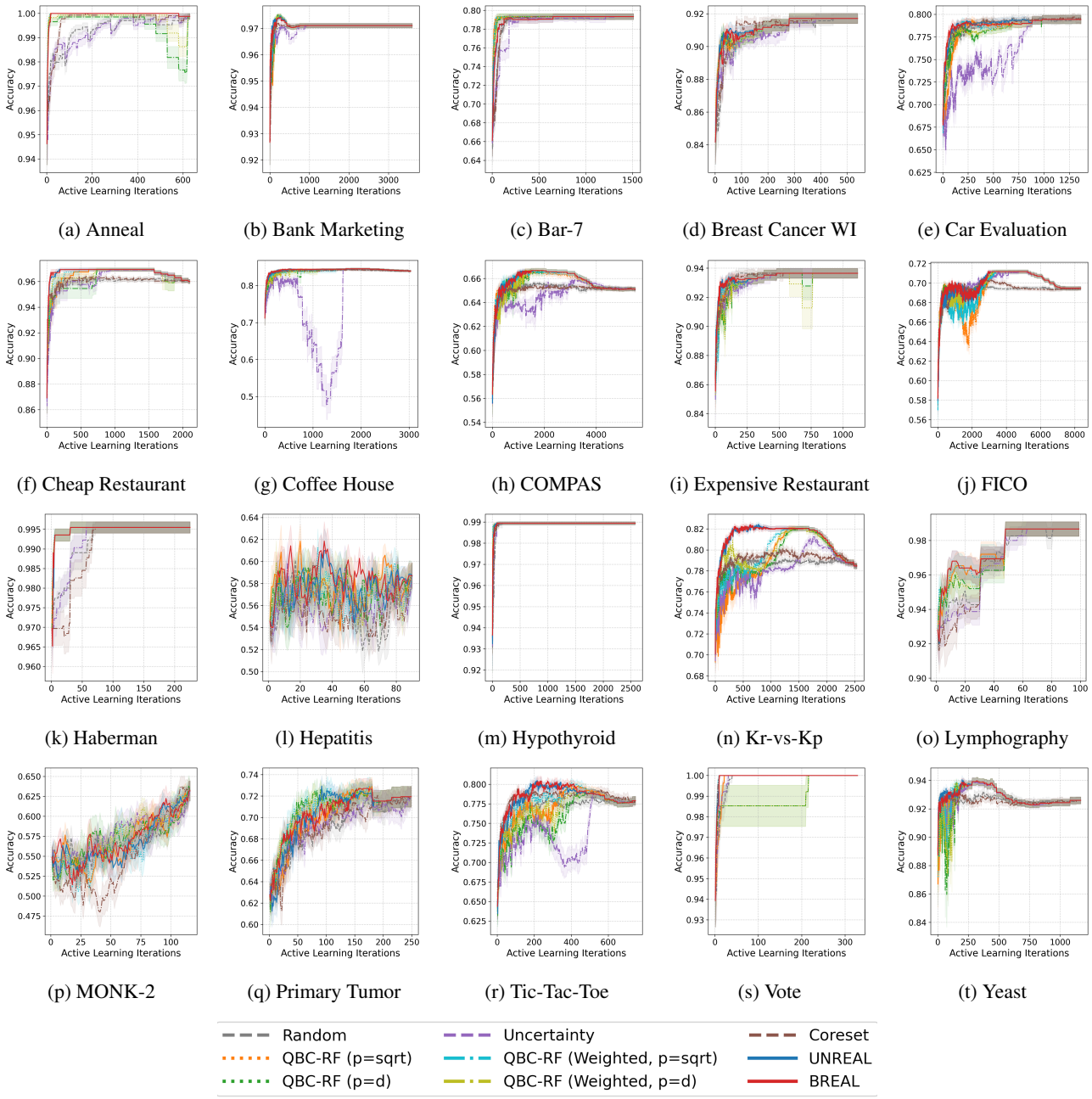
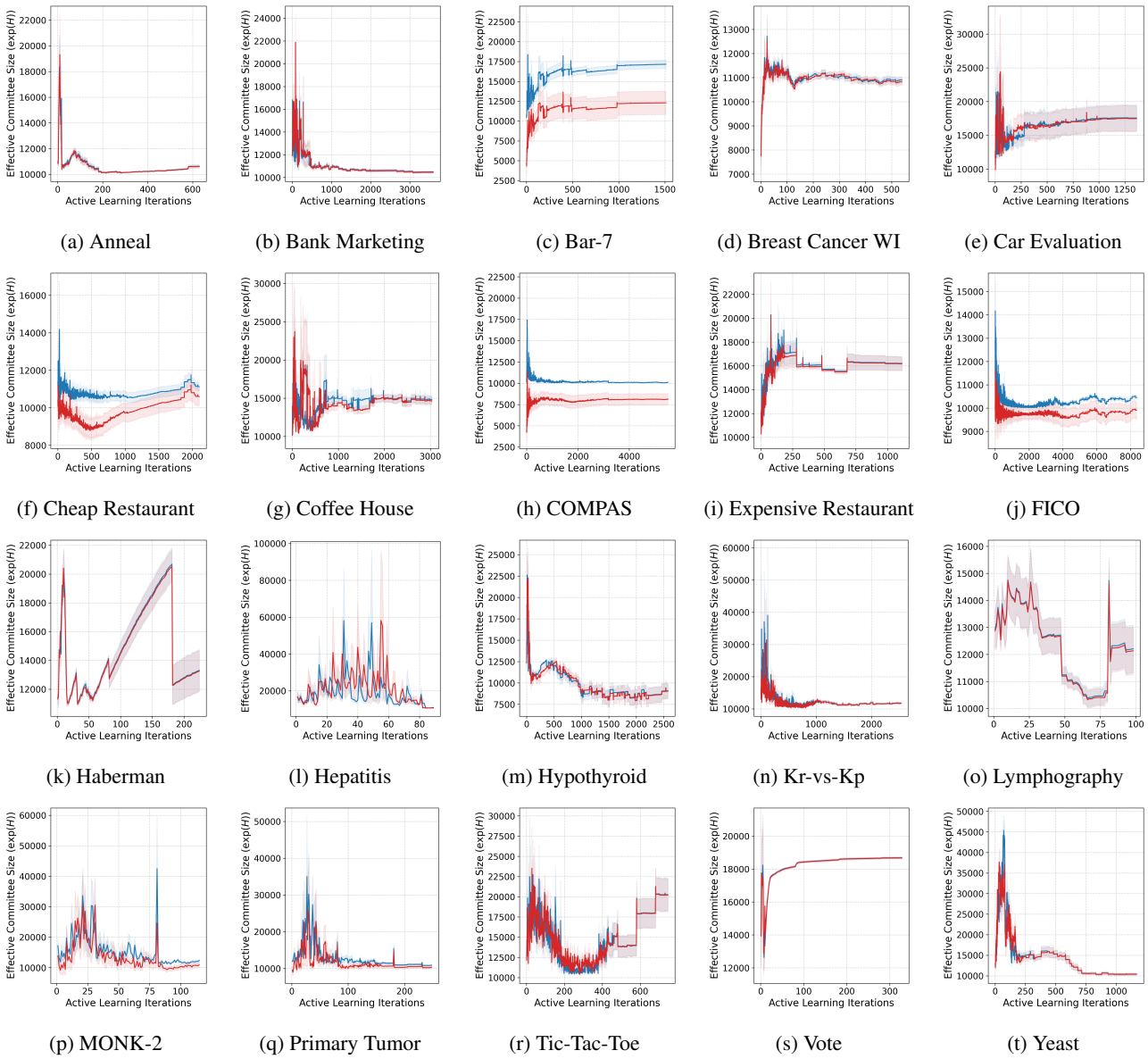


Figure 5: **Active Learning Benchmarks.** Accuracy traces across 20 real-world and structured datasets. UNREAL (Weighted and Uniform) consistently match or exceed baseline performance, demonstrating superior sample efficiency particularly in the early stages of the labeling process.

	Baselines			QBC-RF			Weighted RF		REAL (Proposed)	
Dataset	Random	Uncertainty	Coreset	F=3	F=Sqrt	F=All	F=Sqrt	F=All	UNREAL	BREAL
Synth. Xor Baseline	137.4	80.7	98.0	162.2	164.7	159.4	137.5	140.3	2960.0	3035.0
Synth. Xor Phi 05	168.2	113.2	119.7	196.0	196.0	197.4	176.3	177.1	3014.6	3050.2
Synth. Xor Phi 10	186.3	135.2	135.0	215.9	215.4	217.8	195.6	192.3	3052.7	3043.3
Synth. Xor Phi 25	259.4	203.3	210.6	286.7	284.1	287.1	262.8	266.1	2940.8	2875.1
Synth. Xor Phi 45	294.1	238.5	246.7	318.3	315.3	321.2	293.2	299.3	1674.9	1661.3
Synth. Xor Alpha 50	245.5	195.2	197.1	278.0	277.1	286.5	263.5	265.4	1771.7	1672.2
Synth. Xor Alpha 25	201.8	155.3	152.4	229.6	231.0	243.4	220.9	227.1	3048.2	2955.7
Synth. Xor Alpha 75	249.9	194.7	201.2	288.2	293.9	294.3	270.7	272.7	1567.8	1588.5
Synth. Xor Alpha 100	225.1	176.0	177.6	279.5	281.4	286.6	257.9	263.3	1522.2	1576.3
MEDIAN (Synth.)	**225.1**	**176.0**	**177.6**	**278.0**	**277.1**	**286.5**	**257.9**	**263.3**	**2940.8**	**2875.1**
Anneal	145.3	73.3	83.7	201.0	195.7	196.2	169.3	166.6	1851.0	1536.5
Bank Marketing	1100.6	2417.4	1922.7	3398.2	3361.2	3773.3	3148.9	3423.7	17800.4	13258.9
Bar-7	428.2	218.1	272.3	522.4	522.4	535.4	460.9	474.1	5423.0	5576.8
Breast Cancer Wisconsin	106.3	32.9	33.7	149.1	149.6	151.0	123.6	123.8	1357.8	1358.7
Car Evaluation	506.6	382.0	325.7	653.6	649.5	652.0	576.0	586.4	6155.1	5711.1
Cheap Restaurant	1005.7	1116.7	794.3	1212.0	1221.4	1243.4	1115.6	1102.9	9185.1	8831.7
Coffee House	1998.1	1740.7	1897.1	2507.7	2546.4	2479.2	2328.6	2349.0	22381.6	17902.5
Compas	4557.0	3674.3	5095.2	4985.1	4934.9	5127.7	4357.3	4618.7	27084.6	27142.3
Expensive Restaurant	360.2	232.8	220.3	466.4	467.9	481.9	416.9	423.7	4301.5	4227.2
Fico	25019.3	24959.6	33920.9	24983.4	25006.9	25097.2	23517.2	24598.6	104613.6	69284.3
Haberman	43.1	11.9	12.7	60.7	61.3	61.4	49.3	49.7	612.6	609.4
Hepatitis	122.4	109.8	115.3	135.2	135.4	135.6	127.7	130.2	478.3	488.8
Hypothyroid	650.8	342.0	708.2	986.8	1087.5	1305.1	992.5	1170.0	7461.1	7188.5
Kr-Vs-Kp	19590.7	17263.5	18589.2	18063.6	18646.9	18434.5	17522.9	16722.3	30211.2	29824.1
Lymph	20.3	7.7	8.0	27.7	28.5	29.2	22.5	24.6	235.2	241.2
Monk2	26.6	10.7	11.1	34.9	34.6	35.2	29.4	28.7	393.1	429.3
Primary-Tumor	79.7	46.0	51.3	102.4	101.6	102.9	89.7	89.8	848.5	852.5
Tic-Tac-Toe	355.4	270.1	220.4	434.1	443.6	445.9	396.7	398.9	3329.4	3673.9
Vote	91.4	44.7	52.6	118.1	118.8	119.6	104.6	102.9	1296.3	1326.8
Yeast	5286.4	6831.6	5341.1	7043.2	7226.2	7216.1	7021.3	7245.5	12496.0	12705.6
MEDIAN (Real)	**394.2**	**251.5**	**246.4**	**494.4**	**495.2**	**508.6**	**438.9**	**448.9**	**4862.3**	**4902.0**

Table 3: Median experiment runtime (seconds) across 25 simulation seeds. UNREAL and BREAL denote the standard and Bayesian variants of the UNREAL algorithm, respectively.



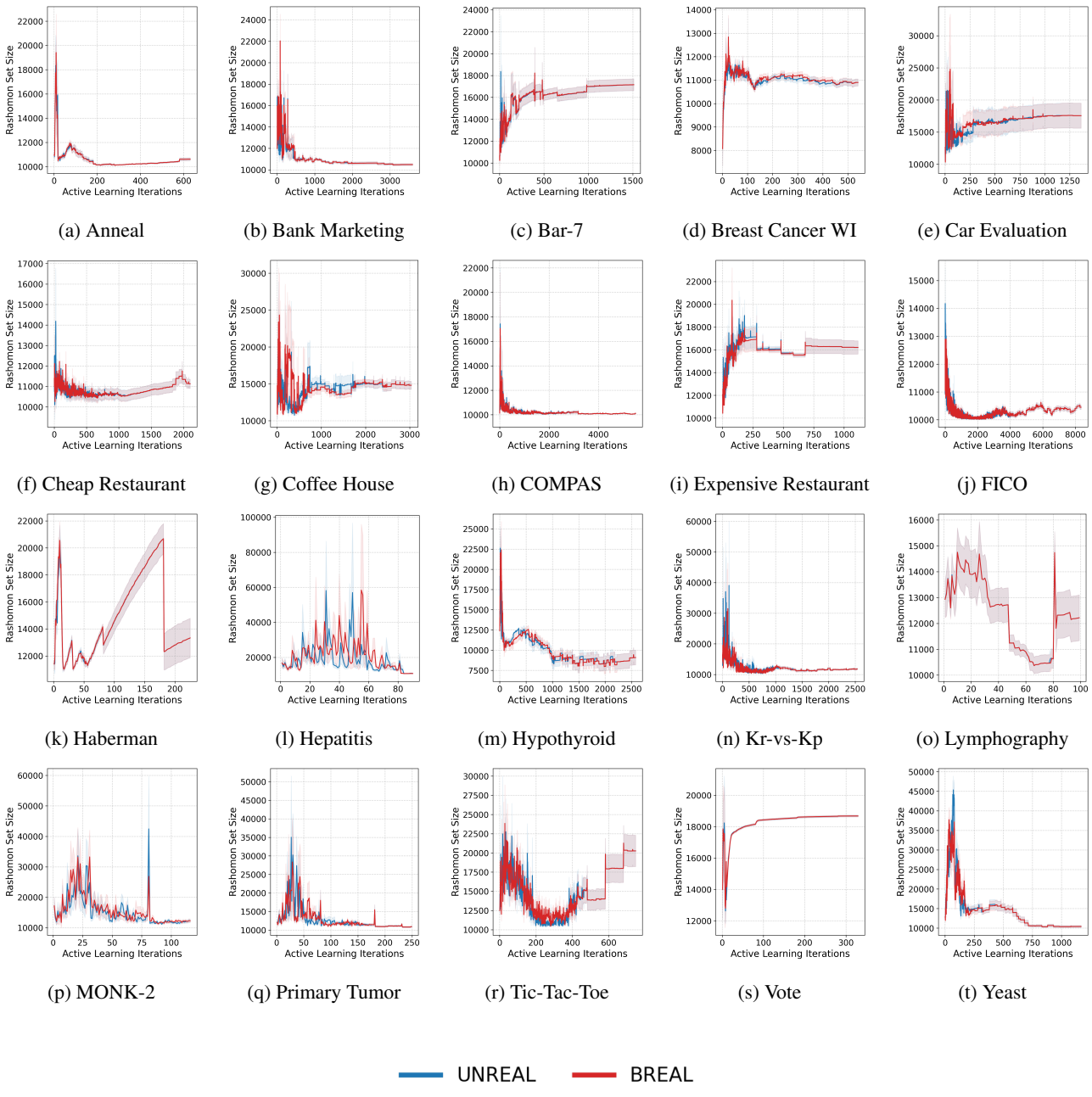
(a) Anneal (b) Bank Marketing (c) Bar-7 (d) Breast Cancer WI (e) Car Evaluation

(f) Cheap Restaurant (g) Coffee House (h) COMPAS (i) Expensive Restaurant (j) FICO

(k) Haberman (l) Hepatitis (m) Hypothyroid (n) Kr-vs-Kp (o) Lymphography

(p) MONK-2 (q) Primary Tumor (r) Tic-Tac-Toe (s) Vote (t) Yeast

Figure 6: **Committee Size History.** Evolution of the Rashomon Effective Committee Size over active learning iterations across 20 benchmark datasets. The ECS serves as a metric of model certainty, quantifying the diversity of plausible theories within the version space as the learner explores the data manifold. A high ECS reflects a broad set of competing hypotheses, while values approaching 1.0 indicate the emergence of a single dominant structural explanation.



(a) Anneal (b) Bank Marketing (c) Bar-7 (d) Breast Cancer WI (e) Car Evaluation

(f) Cheap Restaurant (g) Coffee House (h) COMPAS (i) Expensive Restaurant (j) FICO

(k) Haberman (l) Hepatitis (m) Hypothyroid (n) Kr-vs-Kp (o) Lymphography

(p) MONK-2 (q) Primary Tumor (r) Tic-Tac-Toe (s) Vote (t) Yeast

Figure 7: **Rashomon Set Size Evolution.** The total count of near-optimal models identifying the ϵ-version space during each iteration. These trajectories illustrate how acquiring specific labels prunes the hypothesis space, revealing the sensitivity of model multiplicity to new evidence across diverse dataset geometries.

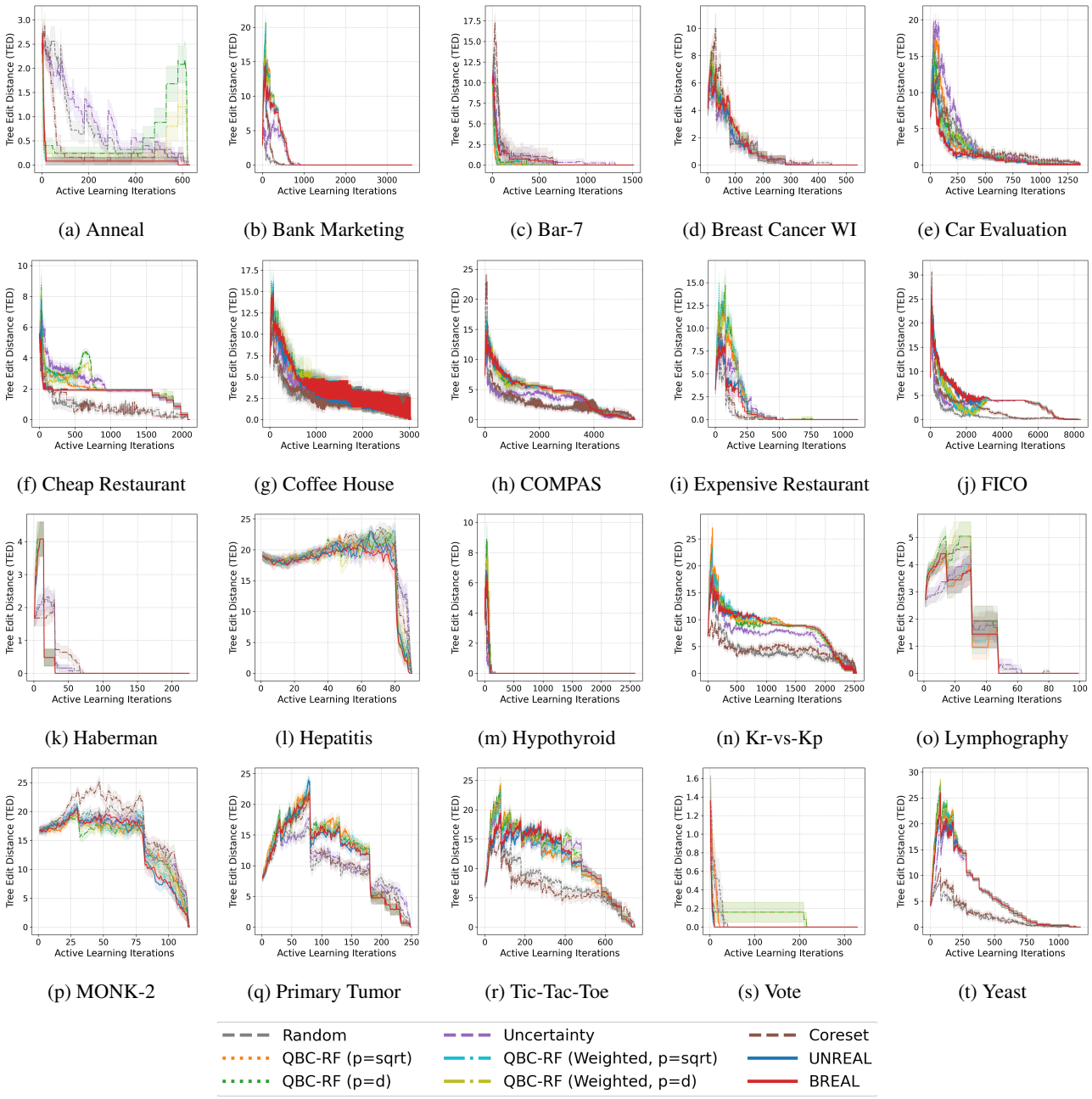


Figure 8: **Tree Edit Distance Benchmarks.** Evaluation of the alignment between the decision logic of the learned hypotheses and the ground-truth Oracle tree. A lower distance indicates that BREAL is not merely achieving predictive success through spurious correlations, but is successfully recovering the true underlying structural logic of the task.

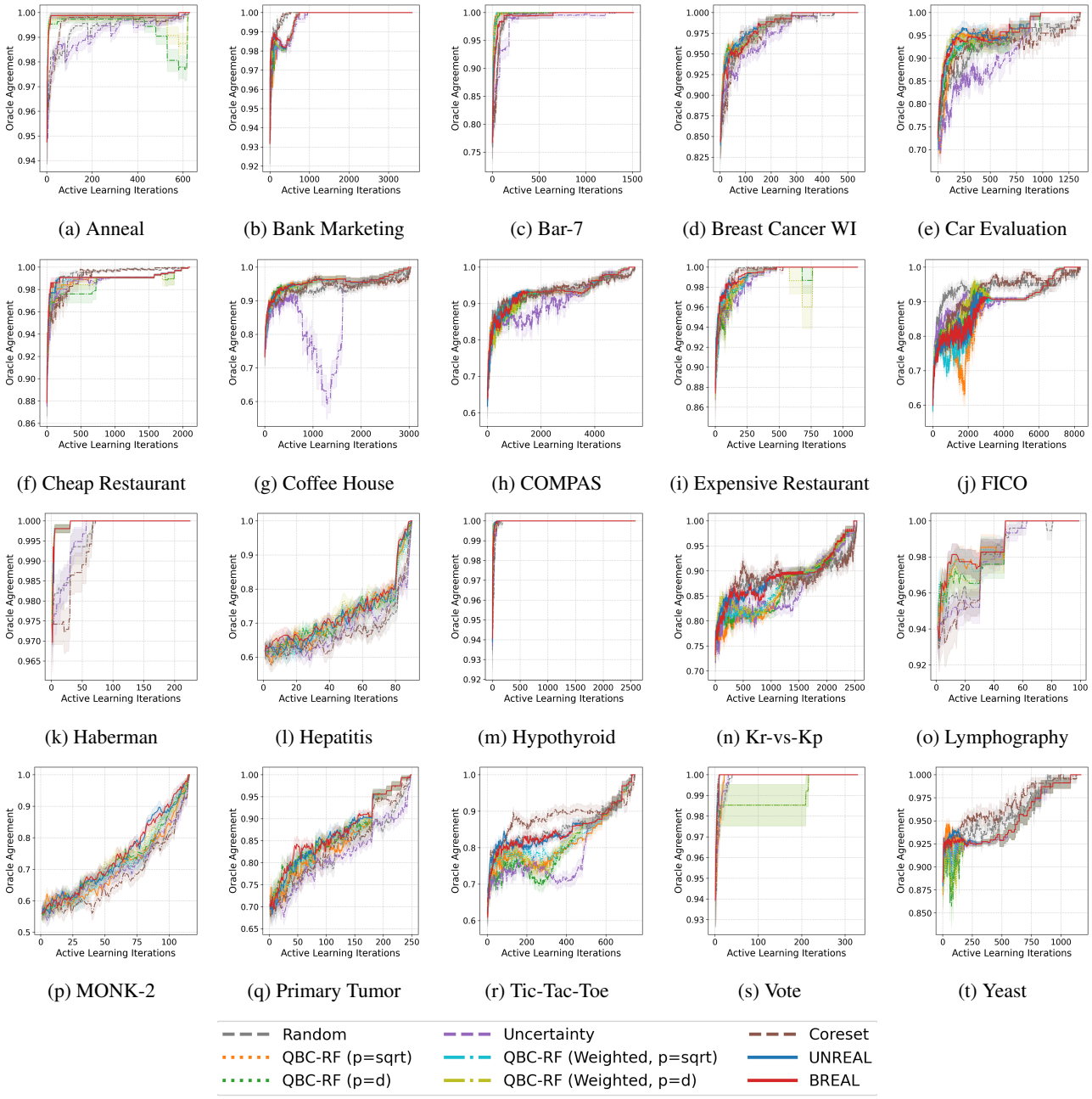
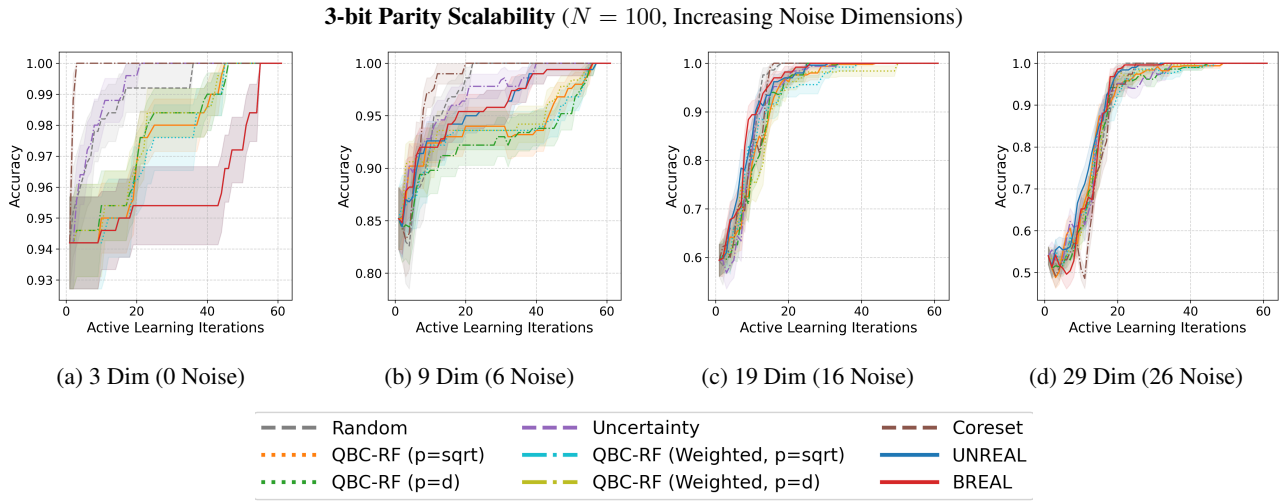


Figure 9: **Oracle Agreement Benchmarks.** Tracking consensus between the active learning committee and a ground-truth "Oracle" (optimal sparse tree fitted to the full dataset). In high-multiplicity environments, Weighted UNREAL achieves higher-fidelity alignment significantly faster than randomized baselines, demonstrating that its structural uncertainty signal successfully identifies the primary decision boundaries defining the task logic.

C EVALUATING REAL ON HIGH-DIMENSIONAL SPARSE PARITY DATASETS

The n-bit parity problem serves as a critical benchmark for evaluating decision trees in high-dimensional sparse scenarios. We utilize a 3-bit parity core (where $Y = 1$ if the sum of the first 3 features is even) and introduce an increasing number of meaningless Bernoulli noise covariates $(0, 6, 16, 26)$. To simulate extreme small-sample constraints, we restrict the total pool to $N = 100$ samples.



(a) 3 Dim (0 Noise) (b) 9 Dim (6 Noise) (c) 19 Dim (16 Noise) (d) 29 Dim (26 Noise)

Figure 10: **Trace Plots on Parity Datasets.** In high-dimensional sparse regimes ($N = 100$, up to 26 noise covariates), all tree-based active learning strategies perform below Passive Sampling as noise-induced patterns dominate. As we increase the number of noisy, useless covariates, QBC-RF becomes increasingly confused by stochastic variance, whereas REAL variants demonstrate superior recovery. By focusing its query strategy on the effective version space, REAL resolves the sparse 3-bit XOR signal more reliably than QBC-RF, eventually matching the accuracy of the other baselines and exceeding the performance of randomized ensembles.

In these experiments, all tree-based active learning methods initially perform worse than fundamental active learning methods Coreset and Uncertainty Sampling and even Passive Sampling. With only 100 samples available and up to 29 total features, the version space is initially dominated by "false" structural patterns induced by noise covariates. This leads to greedy queries that provide less information than a uniform random search of the feature space.

In this low-sample regime, UNREAL and BREAL initially trail their QBC-RF counterparts. We attribute this to the stochastic buffer of randomized ensembles: feature subsetting and bootstrapping allow QBC-RF to smooth over geometric confusion by chance. In contrast, the precision of the REAL framework makes it hypersensitive to local misalignments before the version space has sufficiently collapsed.

As we increase the number of meaningless covariates, a secondary trend emerges: the exhaustive version-space characterization performed by REAL filters out aleatoric noise more effectively than randomized heuristics. While randomized ensembles (QBC-RF) become increasingly confused by the local patterns found in random feature subsets, REAL maintains a global perspective by explicitly identifying all models within the Rashomon set that are consistent with the data.

These Parity results coincide with our findings on model misspecification and label noise in the main manuscript. As the number of noise covariates increases, the Rashomon set expands to include a wider array of potential structural explanations—a phenomenon identical to the expansion seen under label noise (ϕ). By focusing its query strategy on the intersection of these competing hypotheses, REAL ensures that it only labels points that resolve true epistemic ambiguity. This allows the framework to navigate the expanded search space with a level of precision that stochastic approximations cannot match, eventually identifying the 3-bit XOR core once a critical mass of labels is acquired.

D SENSITIVITY ON RASHOMON THRESHOLD

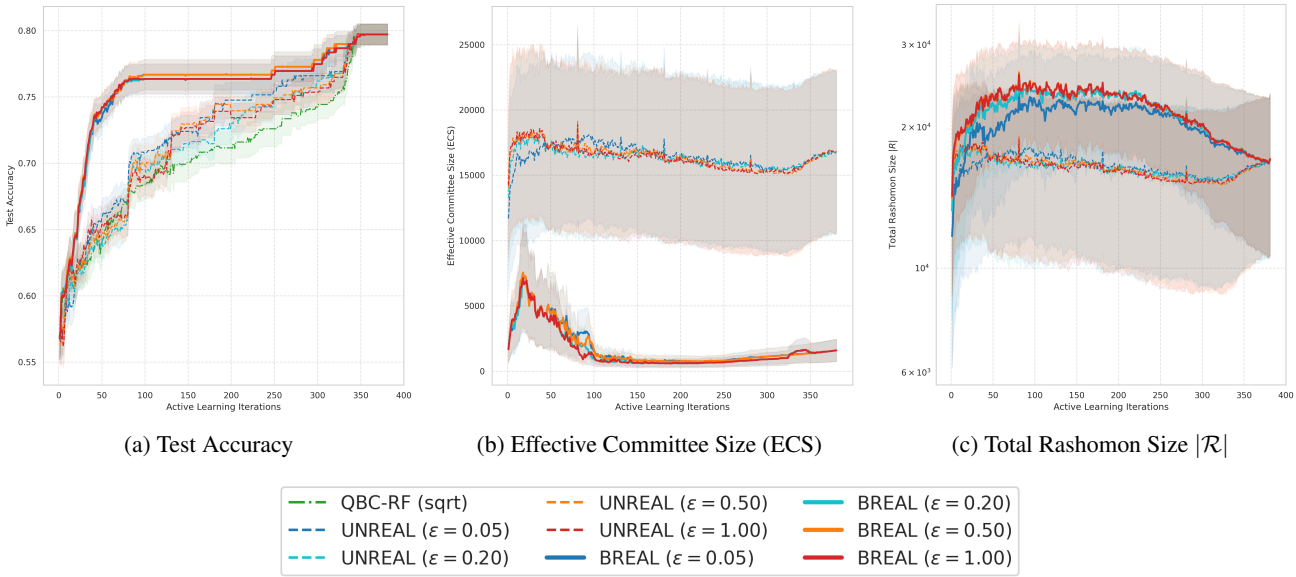
In this section, we present an ablation study evaluating the sensitivity of our framework to the Rashomon threshold, ϵ. This parameter bounds the Rashomon set, dictating the volume of near-optimal models admitted into REAL's committee. A critical concern in version-space approximation is the risk of threshold misspecification: setting ϵ too high risks flooding the committee with suboptimal, functionally redundant hypotheses while setting ϵ too low will limit the diversity of the REAL committee.

Experimental Setup: We evaluated the performance of both BREAL and the unweighted UNREAL baseline across a spectrum of threshold levels, $\epsilon \in \{0.05, 0.20, 0.50, 1.0\}$. The analysis was conducted utilizing a synthetic bimodal parity dataset (`Ablation_Bimodal_Study`), which naturally exhibits high structural multiplicity and complex feature interactions. The DGP is constructed as a mixture of 3-bit and 2-bit parity functions embedded with uninformative noise covariates. This specific configuration guarantees the existence of multiple structurally distinct, near-optimal decision trees, deliberately stressing the framework's ability to filter functional redundancy as the ϵ-boundary expands.

The Version Space Explosion: As anticipated, relaxing the threshold ϵ results in a massive expansion of the enumerated Rashomon set ($|\mathcal{R}|$). As shown in Figure 11(c), the absolute volume of the set quickly reaches upwards of 10^4 models regardless of the threshold. For UNREAL, this influx heavily dilutes the epistemic signal. Figure 11(b) demonstrates that UNREAL's ECS balloons to match this raw multiplicity (exceeding 15,000), causing the vote entropy to become dominated by the sheer volume of redundant structural variants rather than genuine ambiguity.

Robustness via PAC-Bayesian Weighting: The critical finding of this ablation study is the resilience of the BREAL framework under threshold misspecification. Figure 11(b) shows that despite the massive expansion of $|\mathcal{R}|$, BREAL's ECS is aggressively constrained. After a brief initial exploratory spike, the Gibbs weighting forces the ECS to rapidly collapse, stabilizing orders of magnitude lower than its UNREAL counterpart.

This empirical result demonstrates that the PAC-Bayesian Gibbs weighting acts as an elegant and highly effective "structural filter." Even when a heavily relaxed ϵ allows tens of thousands of suboptimal models into the committee, the exponential penalty of the Gibbs posterior naturally mutes their influence. By aggressively down-weighting these peripheral models, BREAL ensures that the active learning signal remains strictly anchored to the high-evidence core of the version space. Consequently, as seen in Figure 11(a), BREAL achieves superior and tightly clustered test accuracy regardless of how loosely the ϵ boundary is set.



(a) Test Accuracy (b) Effective Committee Size (ECS) (c) Total Rashomon Size $|\mathcal{R}|$

Figure 11: **Threshold Sensitivity (ϵ) Ablation.** (a) BREAL maintains tightly clustered, superior accuracy across all ϵ thresholds compared to UNREAL. (b) While UNREAL's ECS explodes to match the raw Rashomon size, BREAL's Gibbs weighting heavily constrains effective committee size, filtering out redundant hypotheses. (c) The raw enumerated Rashomon size ($|\mathcal{R}|$) on a log scale, demonstrating the massive structural multiplicity of the task.