STOP OPTIMIZING FOR THE MIDDLEMAN: RE-ALLOCATING EXPERT TIME TO DIRECTLY IMPROVE SCIENTIFIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Marginal improvements in classifier accuracy on scientific datasets can require significant annotation and optimization effort while yielding little improvement in the downstream quantities scientists ultimately seek to estimate. We study an alternative paradigm: allocate scarce human expertise according to its expected value for downstream scientific inference. Across scientific inference tasks, we show that this strategy can outperform spending the same labeling budget on classifier correction, and analyze differences between what data gets sampled in each case. We also explore classifier fine-tuning within this system and find that it can further boost downstream scientific inference, and that the strongest results come from selecting labels to verify with target-directed review and re-using these labels to update the classifier. Our results all emphasize the need for our field to move beyond classifier accuracy as the central objective when building expert verification systems in AI for science.

1 INTRODUCTION

Scientists are increasingly using ML models to label their data. Predictive models have become remarkably capable across scientific domains (Wang et al., 2023; Vinuesa et al., 2026), but they are often applied to datasets different from those on which they were trained. Under such distribution shifts, predictive performance can degrade (Koh et al., 2021). Human experts must therefore verify or correct model predictions before using them for scientific analyses. However, ***expert time is valuable.*** Scientific datasets can contain millions of observations, and experts can realistically verify only a small fraction of model-generated labels (Vélez et al., 2023; Dumoulin et al., 2025; Kelling et al., 2015). ML has traditionally directed expert review toward improving prediction: adapting models to the target dataset through active learning (Settles, 2009), or routing uncertain predictions for human review through selective prediction (El-Yaniv et al., 2010; Geifman & El-Yaniv, 2017).

Yet in many scientific applications, ***classification is only an intermediate step*** towards downstream scientific inference. An ecologist, for example, may ultimately care about where in the landscape the Black-throated Green Warbler lives, rather than which individual recordings contain its calls. Across scientific disciplines, model-derived labels are similarly aggregated or combined with other measurements for inference. In these settings, predictive improvements matter when they translate into better downstream scientific estimates. But expert effort spent on predictions with little influence on that estimate may have limited scientific value. Recent work has shifted the focus from prediction to the downstream scientific inference it supports (Angelopoulos et al., 2023; Zrnic & Candes, 2024; Haucke et al., 2026b; Hamilton et al., 2026). These works treat model outputs as imperfect measurements, and suggest another way to spend scarce expert effort: ***we could review those expected to be most directly informative about the scientific target itself.***

We study the tradeoff between spending limited expert effort on correcting classification outputs or improving downstream scientific inference directly, given a particular dataset and scientific target. When does adapting a classifier actually improve the scientific estimate? We study these questions across three scientific domains, each represented by a real-world dataset with a distinct modality and three downstream scientific inference tasks (see Figure 1). Our experiments reveal three main findings:

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.

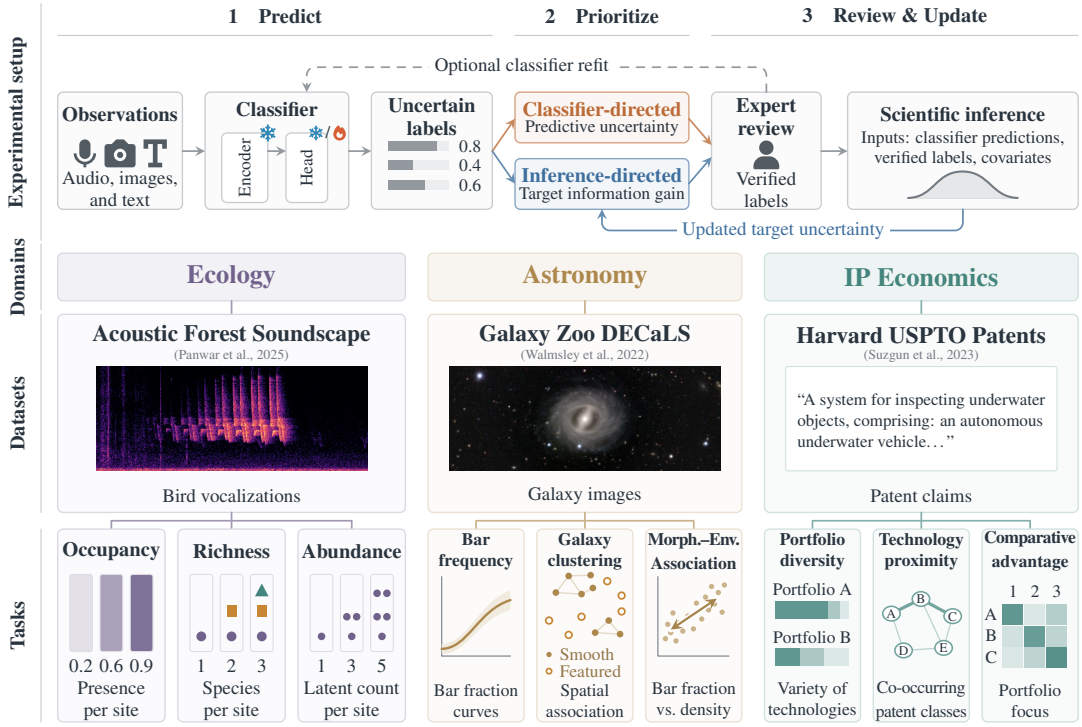

Figure 1: **When the goal is downstream scientific inference and expert labels are scarce, should labeling be classifier- or inference-directed?** Classification models can process large datasets efficiently, but output uncertain and imperfect labels. Expert review is therefore necessary to correct and verify predictions and produce high-quality downstream scientific inference. When expert effort is scarce, how do we allocate this labeling effort effectively?

1. **Classification performance is insufficient to characterize scientific utility.** Models with similar predictive performance can produce substantially different downstream estimates when their predictions are used naively.
2. **Expert review is most effective when allocated toward downstream scientific inference when there is a specific target of interest.** Different scientific tasks prioritize different observations, so expert review should focus on those most relevant to the target.
3. **Target-directed review is consistently the best strategy, even when incorporating classifier fine-tuning.** Better classifiers usually translate to better inference, but standard active learning focused on classifier improvement is a suboptimal learning strategy for downstream scientific inference.

We do *not* argue that classifier performance is irrelevant or that better models are unnecessary. Strong representations and pretrained classifiers still determine the quality of the starting point for downstream analysis. However, our results suggest that classifier accuracy is an imperfect proxy for the scientific target, and scarce expert effort should instead be allocated according to its value for improving downstream inference.

2 Related Work

ML-supported inference in various sciences. Across scientific domains, ML increasingly supports scientific inference by providing intermediate measurements. In ecology, ML-classified camera trap and acoustic observations are integrated with point counts to estimate species richness, activity, occupancy, and abundance (Whytock et al., 2021; Doser et al., 2021; Rhinehart et al., 2022). And in astronomy, predicted galaxy morphologies are assembled into catalogs used to study galaxy populations (Walmsley et al., 2022; Martin et al., 2020). These applications use ML classifications as intermediate measurements, but what remains less clear is how this downstream use should change

the allocation of scarce expert supervision. This includes which predictions should be reviewed, and whether those reviews are most valuable for updating the classifier or the scientific estimate.

Targeted acquisition for prediction and inference. Prior work has emphasized developing sampling strategies that differ primarily in what the acquired labels are meant to improve. Bayesian experimental design selects measurements which maximize expected utility under the current posterior and predictive distribution, often through expected reductions in uncertainty about parameters, predictions, or decisions (Lindley, 1956; Chaloner & Verdinelli, 1995; Ryan et al., 2016; Rainforth et al., 2024). Bayesian active learning instead typically targets the predictive model itself: methods such as BALD select labels expected to reduce uncertainty about model parameters (Houlsby et al., 2011), while active testing and model-selection methods acquire labels to better evaluate or distinguish predictive models (Kossen et al., 2022; Kay et al., 2025).

When predictions are inputs to a downstream estimate, expert review can be directed toward the inference problem itself. Post-prediction inference and related semi-supervised methods treat model outputs as imperfect measurements and use a smaller, fixed set of verified labels to correct or evaluate estimates (Wang et al., 2020; Gronsbell & Cai, 2018). Prediction-powered inference similarly combines predictions with labels to construct valid estimates and confidence intervals (Angelopoulos et al., 2023), while active statistical inference extends this setting to adaptively selected labels (Zrnic & Candes, 2024). Appendix A discusses a broader set of methods that use model predictions or scores for downstream scientific inference. Active measurement goes further by using model-guided sampling while iteratively updating both the predictor and an unbiased estimate of a downstream measurement (Hamilton et al., 2026). More closely aligned with our setting, ACORN applies expected information gain directly to quantities in a species occupancy model, prioritizing review according to expected value for downstream ecological inference (Haucke et al., 2026b). In this paper, we extend this principle of target-directed review to a broader class of scientific estimands and disentangle two roles of expert supervision: how should scarce expert effort be divided between improving predictions and directly reducing uncertainty about the target quantity?

3 Data and methods

We study nine scientific tasks spanning ecological surveys, galaxy morphology catalogues, and patent applications (Figure 1). Each combines uncertain classifications with auxiliary measurements to estimate a population-level quantity. In our active review framework, model predictions guide which observations receive expert verification, and the resulting labels update the scientific estimate and, optionally, the classifier. Across tasks, we vary the review strategy and whether reviewed labels are used for classifier refitting.

3.1 Datasets and scientific tasks

Ecological surveys. Autonomous audio recorders collect repeated observations of wildlife at fixed sites. We use recordings from the Acoustic Forest Soundscape dataset (Panwar et al., 2025). Classifiers predict which species appear in fixed-length clips of these recordings, but ecologists seek higher-level summaries of the populations at those sites. We study three such summaries: *occupancy*, where a species occurs and how its occurrence relates to the environment (MacKenzie et al., 2002), *richness*, how many different focal species are present at each site (Dorazio et al., 2006), and *abundance*, which here corresponds to the latent number of individuals available for detection at a site (Royle & Nichols, 2003). These quantities cannot be derived directly from the classifications. For example, an animal can be present without being detected, and repeated detections can involve the same individual. Statistical models combine repeated observations with environmental information to account for this imperfect observation process. Human review establishes whether a species was detected in a particular observation, helping the model distinguish classifier errors from missed detections.

Galaxy morphology. For galaxy morphology, we use Galaxy Zoo DECaLS (Dark Energy Camera Legacy Survey, Walmsley et al. 2022), which contains galaxy images with community-sourced annotations of their structure. The classifier predicts a binary morphology label from each galaxy image: for the bar frequency task, whether a galaxy contains a stellar bar, and for the clustering task, whether it is smooth or featured. These morphology predictions are then combined with in-

dependently measured galaxy properties to define the downstream scientific targets. *Bar frequency* describes how often galaxies contain an elongated bar of stars across their center, and how this frequency varies with covariates such as color, magnitude, and redshift (Masters et al., 2011). *Spatial clustering* describes whether galaxies with similar appearances tend to occur near one another in space, using coordinate metadata for each image (Skibba et al., 2009). The *morphology-environment association* analysis uses this same per-image coordinate data to ask whether barred galaxies are more common in sparse or dense local environments (Skibba et al., 2012).

Patent applications. The Harvard USPTO Patent Dataset (HUPD, Suzgun et al. 2023) contains patent applications, technology classifications (Cooperative Patent Classification, or CPC), and inventor metadata. *Portfolio diversity* measures how broadly a country's patent portfolio spans technology classes (Petralia et al., 2017). *Technology proximity* measures how often pairs of classes co-occur on a patent relative to their individual frequencies (Breschi et al., 2003). *Comparative advantage* describes which classes are overrepresented in a country's portfolio (Soete, 1987).

Appendix B describes the datasets, scientific targets, and inference models for each task in detail.

3.2 Classifiers and representations

We first embed the underlying audio, images, or text with a frozen encoder, then use these representations to obtain label probabilities for downstream scientific inference (see Figure 1). Specifically, we use Perch 2.0 for audio (van Merriënboer et al., 2026), Zoobot for galaxies (Walmsley et al., 2023), and PatentBERT for CPC classification from a patent's first claim (Lee & Hsiang, 2020), unless noted otherwise. All encoders remain frozen throughout review. For adapting an existing classifier, we learn an additive correction to its pretrained logits. Zoobot was pretrained on Galaxy Zoo data including DECaLS, and our analysis includes examples used in its pretraining. These experiments therefore evaluate downstream use and adaptation of an existing domain-specific classifier, rather than independently held-out classifier generalization. We filter patent filings to those published after PatentBERT's training cutoff (Lee & Hsiang, 2020). Additional implementation details are in Appendix B.

3.3 Review strategies and protocol

At each review round t, a policy assigns each unreviewed candidate i a utility score, representing its estimated value under that policy's objective, and greedily prioritizes candidates with larger utility. Let $\mathcal{D}_t$ denote the information available after round t, and Y_i the answer to candidate review i: species detection, galaxy morphology, or patent–type. The scientific quantity of interest is a target $T = g(\theta)$, where θ represents model parameters and latent states. We use $H[p] = -\mathbb{E}_{X\sim p}[\log p(X)]$ for entropy and $V[p] = \operatorname{tr}\operatorname{Cov}_p(X)$ for total variance.

We compare three families of review strategies: random review, classifier-directed review, and target-directed review. *Random* review samples uniformly without replacement.

Classifier-directed review policies prioritize observations according to uncertainty in the predictive model, without taking into account the downstream scientific target. Let p_{cls} denote distributions conditioned only on classifier inputs and observed labels.

Classifier entropy selects candidates with uncertain predicted labels:

$$\text{Entropy}(i) = H[p_{\text{cls}}(Y_i \mid \mathcal{D}_t)] \tag{1}$$

Classifier BALD instead selects labels expected to reduce uncertainty about the classifier parameters W (Houlsby et al., 2011):

$$\begin{aligned}\text{BALD}(i) &= H[p_{\text{cls}}(Y_i \mid \mathcal{D}_t)] - \mathbb{E}_{W\sim p_{\text{cls}}(W\mid\mathcal{D}_t)}\left[H[p_{\text{cls}}(Y_i \mid W, \mathcal{D}_t)]\right] \\ &= I_{\text{cls}}(Y_i; W \mid \mathcal{D}_t)\end{aligned} \tag{2}$$

Entropy selects examples where the predicted label is uncertain, whereas BALD selects examples whose labels would most reduce uncertainty about the classifier's parameters. As such, we only evaluate BALD when refitting the classifier and not when keeping the classifier fixed.

Target-directed review policies prioritize labels by their expected effect on the scientific target T. We use Target EIG as a general way to align review with downstream inference: it scores each observation by its expected effect on the estimand itself, without requiring a task-specific acquisition heuristic. This is particularly useful when observations contribute unevenly to the scientific target. *Target expected information gain (EIG)* measures the expected decrease in target uncertainty:

$$\mathrm{TargetEIG}(i) = H[p(T \mid \mathcal{D}_t)] - \mathbb{E}_{Y_i \sim p(Y_i \mid \mathcal{D}_t)}[H[p(T \mid \mathcal{D}_t, Y_i)]] \tag{3}$$

Here $p(T \mid \mathcal{D}_t, Y_i)$ is the hypothetical posterior over the scientific target after revealing candidate i's true label, and the expectation averages the possible label outcomes. This ideal utility equals the conditional mutual information $I(Y_i; T \mid \mathcal{D}_t)$. For ecological targets, we instead average marginal information, $|T|^{-1} \sum_j I(Y_i; T_j \mid \mathcal{D}_t)$, using discrete target distributions. Galaxy and patent tasks use approximations described in Appendix C. Appendix E discusses why Active Measurement (Hamilton et al., 2026) and Active Statistical Inference (Zrnic & Candes, 2024) are not included as baselines, and compares Target EIG to Active Statistical Inference on applicable tasks.

Review protocol. Each experiment alternates between inference, selection, and label revelation until the review budget is exhausted. In every round, we greedily select the five unreviewed candidates with the highest utility under the current posterior, reveal their labels, and update the downstream model. In the refit setting, the classifier is also updated with all acquired labels. Unless otherwise noted, we repeat each configuration across five acquisition seeds and report means and standard deviations across seeds. Due to high computational cost, we use a single seed and consequently report no standard deviation for abundance (see Appendix Figure 10). More details on the review protocol can be found in Appendix C.

Evaluation. We evaluate each review strategy by how quickly its downstream scientific estimate approaches a corresponding *oracle* as additional labels are acquired. The oracle is fit using ground-truth labels instead of classifier-provided labels but is otherwise identical to the partially reviewed model. For the ecological tasks, we report site-level mean absolute error (MAE) between posterior mean occupancy, richness, or abundance and the oracle. For the galaxy tasks, we report root mean squared error (RMSE) across the bar-frequency curves, spatial-correlation statistic, or bar–environment association. For patents, we report RMSE across countries' normalized entropies, class-pair log lifts, or country–class log-transformed advantage values. Lower values indicate closer agreement with the oracle. When comparing predictive models directly, we report class-averaged classification accuracy on the underlying labels. The oracles serve as reference estimates for comparing review strategies and should not be interpreted as physical or population-level ground truth.

4 Results and Discussion

4.1 Classification accuracy is an imperfect performance proxy for downstream scientific inference.

Before asking how expert supervision should be allocated, we first explore the relationship between classifier accuracy and scientific estimates. We do this by permuting classifier confidence scores within each ground-truth class 100 times, preserving identical aggregate classifier performance while changing *which* observations receive high or low confidence. We then recompute each downstream estimate with the updated classifier scores. As shown in Figure 2(a), classifications with identical predictive performance can produce substantially different downstream errors across datasets and tasks. The Technology Proximity task on the HUPD data strongly depends on the relative co-occurrence between classes, making it particularly sensitive to synthetic within-class score permutations that do not account for such structure. More broadly, prediction errors can have different effects on the scientific estimate. Systematic errors can matter more or less depending on where they occur and how those observations contribute to the downstream estimand. Thus, two classifiers with the same aggregate accuracy can differ substantially in scientific utility. While deliberately extreme, this controlled experiment emphasizes that aggregate classifier metrics can obscure differences that impact downstream estimates.

We next ask whether the same mismatch appears in the real world across independently trained, state-of-the-art predictive models. Recent bioacoustic models provide a useful test case. Considerable effort has gone into developing increasingly capable pretrained representations for bioacoustics.

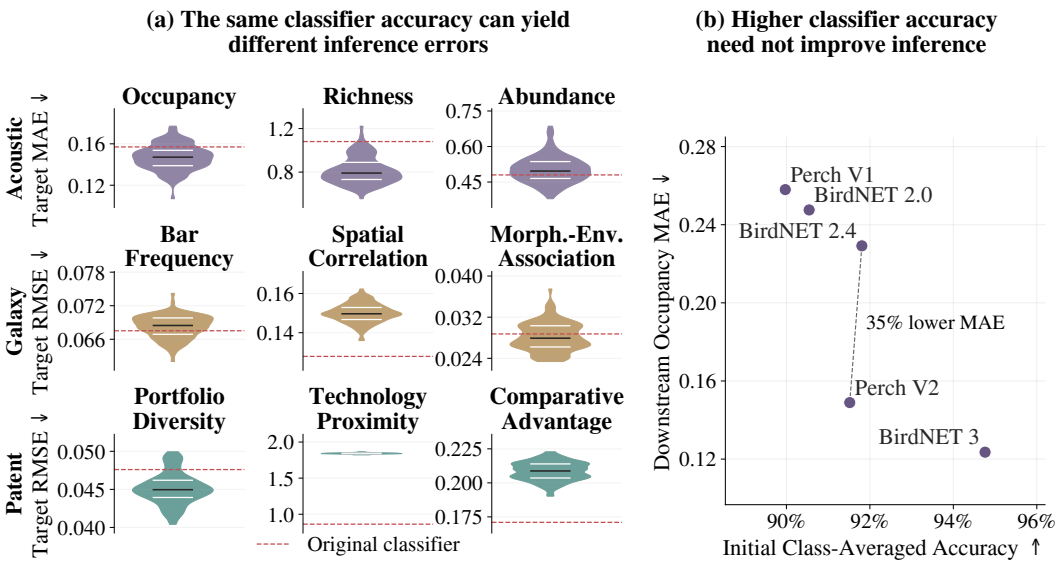

Figure 2: **Classifier accuracy is an imperfect proxy for downstream scientific performance.** (a) We permute classifier score vectors within true classes, leaving classifier accuracy unchanged, and show that even identical classification performance can result in different inference errors. Technology Proximity strongly relies on inter-class co-occurrence and is thus exceptionally affected by the within-class score permutations. (b) The same mismatch occurs across multiple bioacoustic models. Initial class-averaged accuracy does not consistently rank occupancy inference error. For example, BirdNET 2.4 and Perch V2 have similar classification performance, yet Perch V2 produces lower occupancy error. Error scales differ across tasks due to variable dataset and target difficulty.

But if these models ultimately support scientific analyses, do improvements in predictive performance translate into better scientific estimates? Figure 2(b) compares successive models from the BirdNET (Kahl et al., 2021) and Perch V1 (Ghani et al., 2023) and V2 (van Merriënboer et al., 2026) on the same acoustic occupancy task (see Appendix Figure 14 for results on richness and abundance). Better classification performance generally corresponds to lower downstream error, and stronger acoustic models can provide better starting points for downstream scientific inference. However, ***classification accuracy captures only part of what makes a model useful for downstream scientific inference.*** Figure 2(b) shows that BirdNET 2.4 and Perch V2, for example, have similar predictive performance, yet Perch V2 produces 35% lower occupancy error.

Optimizing scientific ML systems with only a focus on aggregate classifier performance can therefore be misaligned with the downstream estimates they ultimately support. This distinction becomes especially important when expert review is limited and it becomes important to identify which observations are most valuable for the downstream quantity itself.

4.2 Target-directed review improves downstream scientific inference

With fixed classifiers, target-directed review generally reduces scientific error more efficiently than random or classifier-entropy review. As Figure 3 shows, Target EIG achieves the lowest mean error at the final review budget and reduces error much more quickly than other strategies in eight out of nine tasks. Because error is measured against the fully reviewed oracle, every strategy must eventually reach zero once all labels are reviewed. Our comparison therefore concerns how efficiently each strategy approaches that endpoint under a limited review budget.

The bar-frequency analysis with the Galaxy Zoo DECaLS dataset is the main exception. All evaluated strategies perform similarly, and the classifier-entropy review can perform better at early budgets before the methods converge. A likely reason is that the dataset is relatively balanced, so there is less benefit from selectively prioritizing rare or especially informative labels.

What drives this improvement over classifier-directed strategies? Table 1 characterizes the samples selected by each strategy across astronomy tasks (see Appendix G.5 for results and anal-

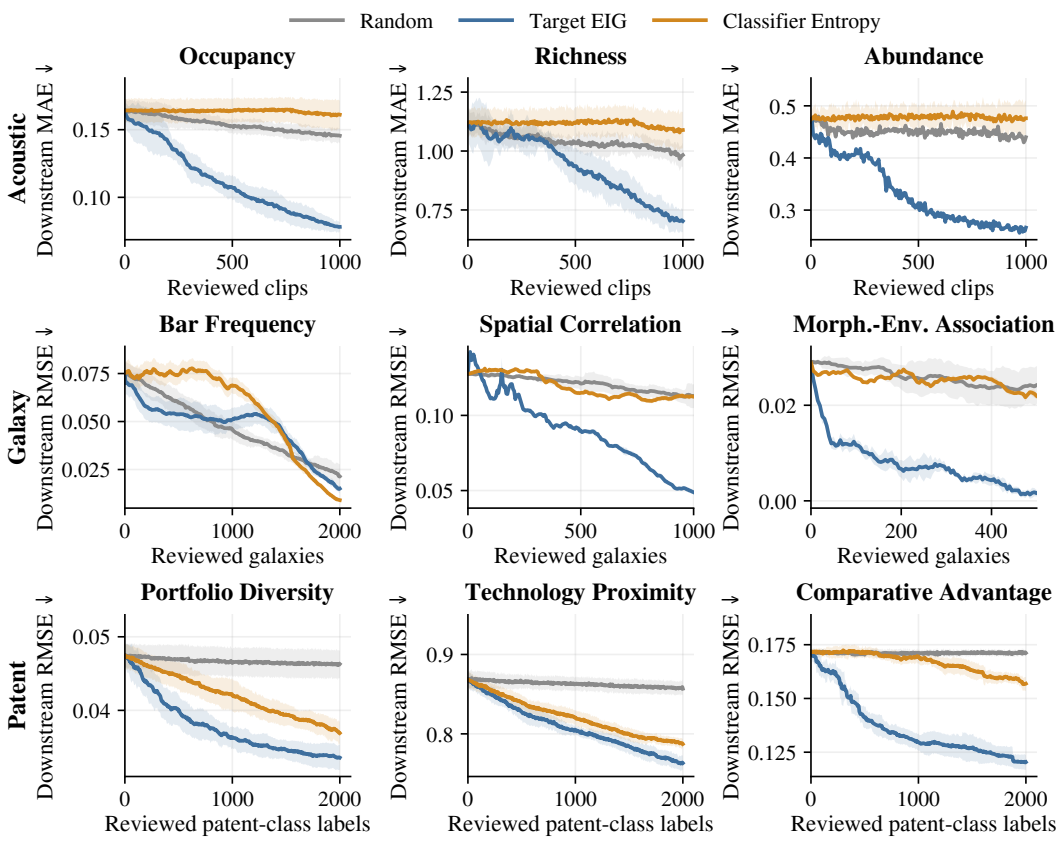

Figure 3: **Target-directed review improves scientific inference with fewer reviews in eight of nine tasks.** Bar frequency is the exception, where all strategies perform similarly. Error relative to the oracle versus number of expert-reviewed labels; lines show means and shading ± 1 SD across seeds. Error scales differ across tasks due to variable dataset and target difficulty.

ysis across acoustic and patent tasks). Target EIG concentrates reviews on the classes, covariate regions, or spatial relationships in the data that are most likely to reduce uncertainty in the downstream estimate. When estimating bar frequency, it preferentially samples galaxies at covariate extremes which have greater influence on the fitted regression. On the spatial correlation task, it instead samples galaxies near previously reviewed examples, increasing the number of informative morphology-labeled pairs that enter the correlation estimate. For quantifying the association between the morphology of a galaxy and its environment, it concentrates reviews in low- and high-density environments, where labels are most informative about how morphology changes across environmental conditions.

These differences arise even though the underlying classification problem is unchanged. Therefore, ***what is useful to label depends on what scientists ultimately want to estimate.*** A galaxy at an extreme covariate value may be valuable for estimating a regression slope but unremarkable for understanding spatial relationships. On the other hand, sampling galaxies within the same neighborhood may strongly constrain a spatial correlation function while contributing little additional information about the morphology of the galaxies themselves. Target-directed acquisition is therefore useful not because it identifies a universally better subset of the data, but because it aligns review with the particular way labels affect the scientific quantity being estimated.

4.3 When does classifier quality matter?

The previous results show that, when expert review is scarce, selecting samples to label based on their estimated downstream scientific value is much more effective than selecting samples to correct model predictions. This does not mean classifier quality is irrelevant. In this section, we explore how

Properties of Selected Data	Downstream Inference Target	Random	Classifier entropy	Target EIG
Covariate extremeness	Bar frequency	0.807 ± 0.004	0.799 ± 0.002	$\mathbf{0.824} \pm 0.002$
	Spatial corr.	0.819 ± 0.009	0.791 ± 0.001	0.709 ± 0.001
	Morph.-Env. Association	0.813 ± 0.028	0.815 ± 0.002	0.842 ± 0.003
Proximity to prior reviews	Bar frequency	7.22 ± 0.06	7.15 ± 0.01	7.12 ± 0.01
	Spatial corr.	5.44 ± 0.11	5.71 ± 0.02	$\mathbf{25.70} \pm 0.10$
	Morph.-Env. Association	3.95 ± 0.11	4.02 ± 0.04	4.42 ± 0.03
Galaxies in low- or high-density environments (%)	Bar frequency	49.84 ± 0.44	49.44 ± 0.07	50.71 ± 0.05
	Spatial corr.	50.00 ± 1.81	50.18 ± 0.10	71.12 ± 0.34
	Morph.-Env. Association	50.60 ± 1.28	47.64 ± 0.27	$\mathbf{99.80} \pm 0.00$

Table 1: **Target-directed review prioritizes task-relevant data.** Entries summarize properties of examples selected over the full review trajectory. Highlighting denotes the property expected to matter most for each target; bold indicates the strategy that emphasizes it most strongly. Values are mean $\pm 95\%$ CI across seeds. See Appendix G.5 for acoustic and patent tasks.

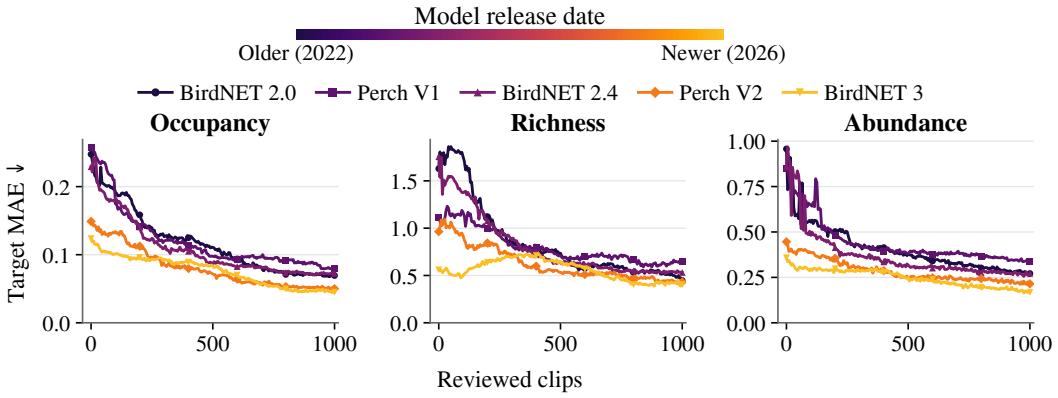

Figure 4: **The starting classifier affects early downstream estimates, but its influence diminishes with increased expert review.** The plots show downstream error under Target-EIG review for five pretrained bioacoustic models on the occupancy, richness, and abundance tasks. Different pretrained classifiers begin from different levels of downstream error, but their performance becomes increasingly similar as reviewed labels accumulate. Color indicates model release date.

improvements in classifier quality—either from newer foundation models or from active learning—relate to downstream scientific inference quality, and further analyze tradeoffs between labeling for classifier *improvement* vs. labeling for improved inference.

How much does the quality of the starting classifier matter? Figure 4 compares several pretrained bioacoustic models under target-directed review. The choice of starting classifier affects downstream scientific inference, particularly early in the review process. As expert labels accumulate, these differences generally narrow, but the relative ordering of classifiers remains similar. ***A stronger starting classifier can provide a better starting point for inference.***

What if you use expert reviews to refit the classifier? So far, we have used expert review to *correct* model predictions. But those same expert labels could instead be selected *update* the classifier itself (i.e., active learning). We compare four ways of using review effort to improve the classifier: (1) only reviewing BALD-selected labels and refitting the classifier; (2) spending the first half of the fixed budget on BALD-selected labels for classifier fitting and the second half on Target-EIG-selected labels for inference; (3) alternating between BALD and Target EIG across review batches;

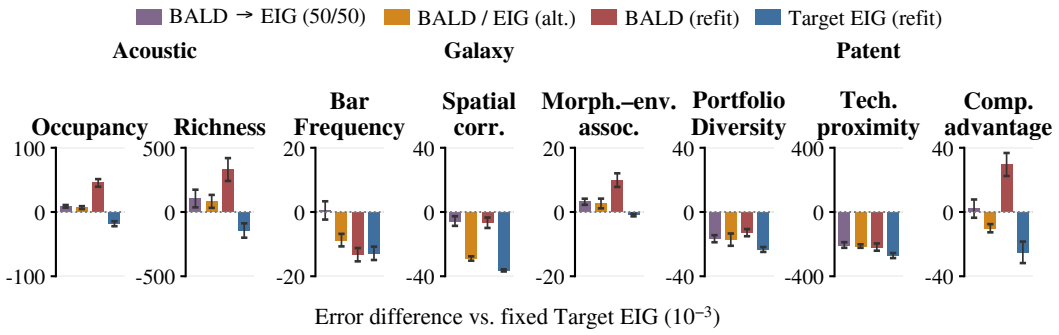

Figure 5: **Refitting helps most when using labels selected for the scientific target.** Bars show the difference in downstream scientific inference error relative to Target EIG with a fixed classifier at each task's review budget. Negative values indicate that refitting improves inference. Error bars show means and ± 1 SD across acquisition seeds. Downstream error is site MAE for acoustic tasks and target RMSE for galaxy and patent tasks. See Appendix Figure 12 for full trajectories.

and (4) only reviewing with Target-EIG, and using labels for classifier refitting *and* downstream scientific inference. We use BALD, not entropy, sampling because it explicitly selects labels expected to reduce uncertainty in the classifier parameters, directly applicable to classifier updating.

Refitting on target EIG-selected labels improves downstream performance, although the size of this benefit varies across tasks. By contrast, allocating part or all of the review budget specifically to BALD does not provide a consistent advantage. Classifier improvement can sometimes help with downstream inference, but it comes at the cost of not reviewing observations with greater direct value to the scientific target. In some cases, the resulting classifier updates can even move the downstream estimate farther away from the oracle. The hybrid strategies provide little evidence that given a specific target, scarce expert verification should be allocated specifically toward classifier-directed acquisition, even with refitting. Appendix G.3 examines this tradeoff in more detail.

We acknowledge using the same adaptively selected labels to both update the predictor *and* as inputs to the downstream analysis violates the underlying statistical assumptions of unbiasedness or calibrated uncertainty (Farquhar et al., 2021). Related work provides such guarantees in more structured settings: Hamilton et al. (2026) updates the predictive model while using importance sampling to retain unbiased estimates of finite-population measurements. Our experiments instead evaluate error on the fixed scientific target of interest directly.

Classifier improvement still matters. Better predictions can improve downstream inference for the many observations that remain unreviewed. The challenge is to capture those benefits without spending scarce expert effort at the expense of the scientific target. Future work should therefore jointly optimize classifier improvement and downstream inference. For example, when two observations are similarly informative about the downstream target, acquisition could prefer the one that also provides greater value for improving the classifier. Developing these strategies, with guarantees for adaptively reusing labels for both prediction and inference, could capture the benefits of classifier adaptation without sacrificing downstream benefits of target-directed review.

5 Conclusion

Scientific ML should optimize expert verification for downstream scientific inference, not just to correct predictive outputs. Which errors your model makes matters, and classification accuracy does not necessarily correspond to performance on downstream scientific inference. We find that scarce expert effort is used most effectively when reviews are selected for their expected value to the downstream target. Even when reviewed labels are used to refit the classifier, target-directed selection provides the most reliable gains. Improving predictions alone is therefore not enough. Expert verification should be aligned with the scientific quantities we ultimately want to estimate.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research methodology or experiments, implementing methods, and cleaning and reformatting datasets. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, supporting qualitative and thematic data analysis, and interpreting results. Generating synthetic data sets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for creating and modifying scientific figures or images, creating and editing software code, drafting parts of the text, brainstorming, searching for information, and identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code and other LLM-generated outputs were verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide code and documentation to reproduce our results, including figures and tables, at `https://anonymous.4open.science/r/bca-iclr-supplemental-B52A`.Appendix B includes additional details on datasets and tasks. Appendix D includes additional implementation details on model fitting.

REFERENCES

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. Quantification via probability estimators. In *2010 IEEE International conference on data mining*, pp. 737–742. IEEE, 2010.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/jax-ml/jax`.

Stefano Breschi, Francesco Lissoni, and Franco Malerba. Knowledge-relatedness in firm technological diversification. *Research policy*, 32(1):69–87, 2003.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Adira Cohen, Erin M Schliep, Roland Kays, Mohammad Alyetama, and Matthew Snider. Improving ecological inference and uncertainty quantification from camera trap data through the fusion of ai confidences and manual annotations. *arXiv preprint arXiv:2605.13660*, 2026.

Robert M Dorazio, J Andrew Royle, Bo Söderström, and Anders Glimskär. Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4): 842–854, 2006.

Jeffrey W. Doser, Andrew O. Finley, Aaron S. Weed, and Elise F. Zipkin. Integrating automated acoustic vocalization data and point count surveys for estimation of bird abundance. *Methods in Ecology and Evolution*, 12(6):1040–1049, 2021. doi: https://doi.org/10.1111/2041-210X.13578. URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13578`.

Vincent Dumoulin, Otilia Stretcu, Jenny Hamer, Lauren Harrell, Rob Laber, Hugo Larochelle, Bart van Merriënboer, Amanda Navine, Patrick Hart, Ben Williams, et al. The search for squawk: Agile modeling in bioacoustics. *arXiv preprint arXiv:2505.03071*, 2025.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *arXiv preprint arXiv:2101.11665*, 2021.

George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):579–594, 2018.

Max Hamilton, Jinlin Lai, Wenlong Zhao, Subhransu Maji, and Daniel Sheldon. Active measurement: Efficient estimation at scale. *Advances in Neural Information Processing Systems*, 38: 145769–145807, 2026.

Timm Haucke, Lauren Harrell, and Sara Beery. Biolith: Bayesian ecological modeling in python, March 2026a. URL https://doi.org/10.5281/zenodo.19358649.

Timm Haucke, Lauren Harrell, Justin Kay, Mary Clapp, and Sara Beery. Targeted review for ai-assisted biodiversity surveys: Active continuous-score occupancy modeling, 2026b. URL https://arxiv.org/abs/2609.25657.

Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

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

Stefan Kahl, Connor M Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4594–4604, 2025.

Steve Kelling, Daniel Fink, Frank A La Sorte, Alison Johnston, Nicholas E Bruns, and Wesley M Hochachka. Taking a 'big data'approach to data quality in a citizen science project. *Ambio*, 44 (Suppl 4):601–611, 2015.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Thomas Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *Advances in Neural Information Processing Systems*, 35:24557–24570, 2022.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 18(14):1–45, 2017.

Jieh-Sheng Lee and Jieh Hsiang. Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965, 2020.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Darryl I MacKenzie, James D Nichols, Gideon B Lachman, Sam Droege, J Andrew Royle, and Catherine A Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

Garreth Martin, Sugata Kaviraj, Alex Hocking, Shaun C Read, and James E Geach. Galaxy morphological classification in deep-wide surveys via unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society*, 491(1):1408–1426, 2020.

Karen L Masters, Robert C Nichol, Ben Hoyle, Chris Lintott, Steven P Bamford, Edward M Edmondson, Lucy Fortson, William C Keel, Kevin Schawinski, Arfon M Smith, et al. Galaxy zoo: bars in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 411(3):2026–2034, 2011.

Pooja Panwar, Wyatt J. Cummings, Sharon Martinson, David A. Lutz, Hannah ter Hofstede, Aaron Weed, Matthew P. Ayres, and Laurel B. Symes. Large-scale, stratified, fully annotated acoustic forest soundscape dataset of avian vocalizations from eastern north america, 2025. URL https://zenodo.org/doi/10.5281/zenodo.18041380.

Sergio Petralia, Pierre-Alexandre Balland, and Andrea Morrison. Climbing the ladder of technological development. *Research Policy*, 46(5):956–969, 2017.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Tessa A Rhinehart, Daniel Turek, and Justin Kitzes. A continuous-score occupancy model that incorporates uncertain machine learning output from autonomous biodiversity surveys. *Methods in Ecology and Evolution*, 13(8):1778–1789, 2022.

J Andrew Royle and James D Nichols. Estimating abundance from repeated presence–absence data or point counts. *Ecology*, 84(3):777–790, 2003.

Elizabeth G Ryan, Christopher C Drovandi, James M McGree, and Anthony N Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.

Burr Settles. Active learning literature survey. 2009.

Ramin A Skibba, Steven P Bamford, Robert C Nichol, Chris J Lintott, Dan Andreescu, Edward M Edmondson, Phil Murray, M Jordan Raddick, Kevin Schawinski, Anže Slosar, et al. Galaxy zoo: disentangling the environmental dependence of morphology and colour. *Monthly Notices of the Royal Astronomical Society*, 399(2):966–982, 2009.

Ramin A Skibba, Karen L Masters, Robert C Nichol, Idit Zehavi, Ben Hoyle, Edward M Edmondson, Steven P Bamford, Carolin N Cardamone, William C Keel, Chris Lintott, et al. Galaxy zoo: the environmental dependence of bars and bulges in disc galaxies. *Monthly Notices of the Royal Astronomical Society*, 423(2):1485–1502, 2012.

Luc Soete. The impact of technological innovation on international trade patterns: the evidence reconsidered. *Research policy*, 16(2-4):101–130, 1987.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem Sarkar, Scott D Kominers, and Stuart Shieber. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *Advances in neural information processing systems*, 36:57908–57946, 2023.

Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics, 2026. URL https://arxiv.org/abs/2508.04665.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved r$\hat{}$for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718, 2021.

Juliana Vélez, William McShea, Hila Shamon, Paula J Castiblanco-Camacho, Michael A Tabak, Carl Chalmers, Paul Fergus, and John Fieberg. An evaluation of platforms for processing camera-trap data using artificial intelligence. *Methods in Ecology and Evolution*, 14(2):459–477, 2023.

Ricardo Vinuesa, Paola Cinnella, Jean Rabault, Hossein Azizpour, Stefan Bauer, Bingni W Brunton, Arne Elofsson, Elias Jarlebring, Hedvig Kjellström, Stefano Markidis, et al. Decoding complexity through machine learning is redefining scientific discovery. *Communications Physics*, 9(1):168, 2026.

Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, Kyle W Willett, Steven Bamford, Lee S Kelvin, Lucy Fortson, Yarin Gal, et al. Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2022.

Mike Walmsley, Campbell Allen, Ben Aussel, Micah Bowles, Kasia Gregorowicz, Inigo Val Slijepcevic, Chris J Lintott, Anna M Scaife, Maja Jabłońska, Kosio Karchev, et al. Zoobot: Adaptable deep learning models for galaxy morphology. *Journal of Open Source Software*, 8(85), 2023.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117 (48):30266–30275, 2020. doi: 10.1073/pnas.2001238117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2001238117.

Robin C Whytock, Jędrzej Świeżewski, Joeri A Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el din, Kelly Boekee, Stephanie Brittain, Anabelle W Cardoso, et al. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12(6):1080–1092, 2021.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Forty-first International Conference on Machine Learning*, 2024.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

A METHODS THAT USE ML SCORES FOR INFERENCE

The idea of using ML outputs as imperfect measurements for downstream statistical inference is not new: several methods explicitly account for prediction error while still extracting useful information from model outputs. Methods focused on quantification estimate class prevalence from classifier outputs, including probabilistic variants that aggregate or correct soft prediction scores to account for classification error (Forman, 2008; Bella et al., 2010). Post-prediction inference instead corrects downstream statistical analyses that use predicted outputs (Wang et al., 2020), while prediction-powered inference and cross-prediction combine many model predictions with smaller labeled samples for more statistically valid and efficient estimates with confidence intervals (Angelopoulos et al., 2023; Zrnic & Candès, 2024). More structured approaches model prediction scores directly. For example, Cohen et al. (2026) combine camera-trap confidence vectors with manual annotations in a Bayesian hierarchical model for ecological inference. In this paper, we ask which observations should receive scarce expert labels in the first place.

B DATASETS AND DOWNSTREAM SCIENTIFIC INFERENCE TASKS

This appendix defines the observations, labels, and downstream targets used in each domain. An observed label describes what can be identified in a recording, image, or text segment. A scientific target describes a property of the population that generated those observations. The distinction is particularly consequential in ecology, where even a correctly identified negative observation does not establish that a species is absent from the site.

B.1 ECOLOGICAL DATASETS AND REVIEW UNITS

Acoustic Forest Soundscape. We use the ACAD, MABI, and SIMR subcollections of the annotated forest soundscape dataset from eastern North America (Panwar et al., 2025). We treat each recording location as a site, and each ten-minute recording as a repeated observation. A species label is positive when that species appears in the recording's annotation file. One review reveals one species–recording binary label. Site predictors (related to how likely a species is to be occupying a site) are tree-canopy cover, canopy height, and elevation. Detection predictors (modeling how likely a species is to be detected if present) are time since sunrise, day of year, and categorical subcollection (ACAD, MABI, and SIMR) indicators.

Continuous ecological covariates are standardized, while binary indicators retain their original coding. Missing covariates are imputed using the median for continuous variables or the mode for indicators where observations are available. All three acoustic tasks use the same subset of ten focal species: AMCR, BHVI, BRCR, BTNW, MAWA, PUFI, RESQ, WIWR, WTSP, and YBSA (see Appendix Figure 6 for the distribution across the classes). These were sampled randomly from the 50 species most common in the dataset and supported by Perch V2 (van Merriënboer et al., 2026), thereby excluding rare species with very few observations. The final dataset contains 104 sites and 1,302 recordings, giving 13,020 species–recording review units. Species share a review budget and are not treated as independent experimental replicates. The classifier-history comparison instead uses the nine species supported by every model, excluding RESQ.

B.2 ECOLOGICAL OBSERVATION MODEL

For one species, let $Y_{jk} \in \{0, 1\}$ indicate detection in observation k at site j, and let u_{jk} be its classifier score. The ecological models below specify a detection probability q_{jk} conditional on the latent site state. Following continuous-score occupancy modeling (Rhinehart et al., 2022) and ACORN (Haucke et al., 2026b), we model score distributions conditional on the true observation label:

$$u_{jk} \mid Y_{jk} = y, \eta \sim \mathcal{N}(\mu_y, \sigma^2), \qquad y \in \{0, 1\}, \tag{4}$$

where η denotes species-specific calibration parameters, with ordered class means and a shared within-species scale. Scores are centered and scaled using the initial head predictions. The acoustic results use a joint model for emission calibration and ecological inference, with community priors shared across species. Conditional on calibration parameters, let $f_y(u)$ denote the corresponding

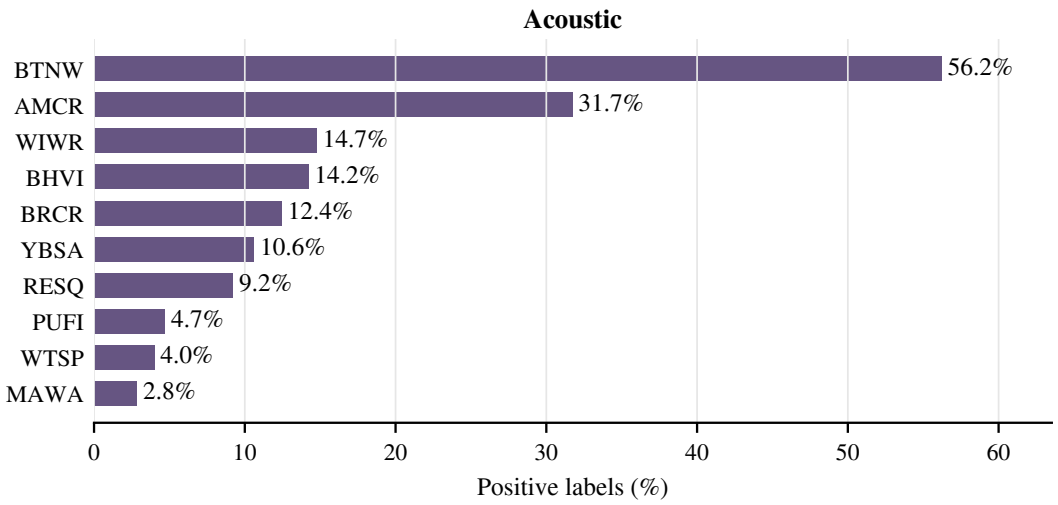

Figure 6: **Distribution of classes in the Acoustic Forest Soundscape dataset.**

class-conditional density. An unreviewed score contributes

$$L_{jk} = (1 - q_{jk}) f_0(u_{jk}) + q_{jk} f_1(u_{jk}) \;\propto\; (1 - q_{jk}) + q_{jk} \frac{f_1(u_{jk})}{f_0(u_{jk})} \tag{5}$$

to the ecological likelihood. The density ratio expresses how strongly the score supports a detection. Once reviewed, the observation contributes $q_{jk}^{Y_{jk}}(1 - q_{jk})^{1-Y_{jk}} f_{Y_{jk}}(u_{jk})$, retaining the score likelihood for calibration. Reviewed labels are treated as correct within the model, although a negative label still allows the species to be present but undetected.

All acoustic fits, including the oracles, use variational inference (see Appendix D for details). In the fixed setting, the embedding head remains fixed and emission parameters are frozen at their initial posterior means. In the refit setting, the head and joint emission calibration incorporate newly reviewed labels. Ecological inference uses every acquired label in both cases.

B.2.1 Occupancy

Occupancy separates whether a species is present from whether it is observed (MacKenzie et al., 2002). Let Z_j denote its latent presence at site j, $\mathbf{x}_j$ the site predictors, and $\mathbf{v}_{jk}$ the observation predictors, with both predictor vectors including an intercept. The model is

$$Z_j \sim \text{Bernoulli}(\psi_j), \qquad \text{logit}(\psi_j) = \mathbf{x}_j^\top \boldsymbol{\beta}, \tag{6}$$

$$Y_{jk} \mid Z_j \sim \text{Bernoulli}(Z_j p_{jk}), \qquad \text{logit}(p_{jk}) = \mathbf{v}_{jk}^\top \boldsymbol{\alpha}. \tag{7}$$

Thus $q_{jk} = Z_j p_{jk}$ in Appendix B.2. Repeated observations help distinguish an unoccupied site from an occupied site where the species went undetected during the survey. The model assumes that site presence is constant across the included observations and that observations are independent conditional on presence and the regression parameters. For each coefficient, a Gaussian community prior across the ten species has covariance $0.6^2 I + 0.7^2 \mathbf{1}\mathbf{1}^\top$. Occurrence and detection intercept means are 0 and -0.5, respectively. Slope means are zero, with slope draws additionally scaled by 0.3. The same ecological coefficient priors are used for richness and abundance.

The occupancy task targets the latent presence indicators $\{Z_{js}\}$ across species and sites. Evaluation compares their posterior means, $\Pr(Z_{js} = 1 \mid \mathcal{D}_t)$, instead of habitat-only regression probabilities ψ_{js} or regression coefficients. Shared coefficients allow a review at one site to change estimates at other sites. The oracle fits the corresponding occupancy model to all observation labels, with the same covariate specification.

B.2.2 RICHNESS

Richness concerns diversity within the specified species panel $\mathcal{S}$ (Dorazio et al., 2006). We use the same joint community occupancy model and species panel as the occupancy task, changing the scientific target and acquisition utility. The target is the vector of site richness values,

$$R_j = \sum_{s \in \mathcal{S}} Z_{js}, \qquad \mathbb{E}[R_j \mid \mathcal{D}_t] = \sum_{s \in \mathcal{S}} \Pr(Z_{js} = 1 \mid \mathcal{D}_t). \tag{8}$$

These probabilities condition on the site's observations as well as the inferred environmental relationships. They therefore differ from simply summing the regression probabilities ψ_{js}. Conditional on a joint coefficient draw, convolution of the species presence distributions gives a Poisson-binomial richness distribution. Averaging over joint draws retains association induced by the shared community prior and site coefficients.

The target counts species present, including those that may have gone undetected. The number of species actually detected is a lower bound under the assumed error-free ground truth labels. Primary evaluation compares posterior site means with the same latent richness model fitted to all labels. It measures recovery of an oracle model, and does not establish the true richness of unobserved species outside the focal subset of species.

B.2.3 ABUNDANCE

Abundance uses a Royle–Nichols model to infer latent counts from repeated binary detections (Royle & Nichols, 2003). The mechanism is that more individuals provide more opportunities to detect at least one. For a single species,

$$N_j \sim \text{Poisson}(\lambda_j), \qquad \log \lambda_j = \mathbf{x}_j^\top \boldsymbol{\beta}, \tag{9}$$

$$Y_{jk} \mid N_j \sim \text{Bernoulli}\big(1 - (1 - p_{jk})^{N_j}\big), \qquad \text{logit}(p_{jk}) = \mathbf{v}_{jk}^\top \boldsymbol{\alpha}. \tag{10}$$

Here p_{jk} is the detection probability for an individual, and $q_{jk} = 1 - (1 - p_{jk})^{N_j}$ is the probability of observing at least one individual. We target the latent site counts $\{N_j\}$, reporting their posterior means and uncertainty, instead of the Poisson rates $\{\lambda_j\}$ or the number of positive observations. The count distribution has unbounded support.

The model assumes a closed site population across the included observations, independent individual detections with a common probability within an observation, and conditional independence across observations. Abundance and detection can be difficult to distinguish when detection histories are sparse or saturated. Log-rate and detection coefficients use the community priors described for occupancy. We marginalize the latent counts during fitting and use adaptive tail-controlled integration for predictions, with tolerance 10^{-8} and no fixed biological count limit. These assumptions and priors are part of the abundance estimand's model dependence. The oracle uses all binary labels in this same model. Neither the reference nor a reviewed label supplies *directly* measured site abundance.

B.3 GALAXY ZOO DECALS DATA AND MORPHOLOGY LABELS

All three galaxy tasks use Galaxy Zoo DECaLS images and GZD-5 volunteer morphology annotations (Walmsley et al., 2022). Our prepared catalogue contains 228,059 galaxies from the training and test sets, with ancillary SDSS photometry and redshifts matched within two arcseconds. We use raw vote fractions, require at least ten votes on the smooth-or-featured question and an artifact vote fraction no greater than 0.2, and assign a positive review label when the relevant vote fraction is at least 0.5. For bar frequency and environment association, the positive fraction is the sum of strong- and weak-bar votes. For clustering, the positive class is smooth and the complement is treated as featured.

These labels assess visible structure, not error-free physical truth. Review cannot recover features unresolved in the images. We generate 512 completed morphology catalogues per round. Bar frequency and environmental association sample classifier coefficients and then unreviewed labels conditional on those coefficients. Spatial clustering instead samples independent Bernoulli labels at posterior mean classifier probabilities, excluding head-parameter uncertainty from the scientific target. Reviewed and initialization labels remain fixed in every completion. Population analyses do not feed back into the image classifier.

Unless otherwise noted, each configuration uses five acquisition seeds and records trajectories through 2,000 additional reviews for bar frequency, 1,000 for spatial correlation, and 500 for morphology-environment association (see Figure 3).

B.3.1 BAR FREQUENCY AND GALAXY PROPERTIES

The bar analysis, based on (Masters et al., 2011), selects face-on disk galaxies using featured and face-on vote fractions of at least 0.5, with at least ten bar-question votes. We require $0.01 < z < 0.06$, absolute r-band magnitude $M_r < -19.38$, and finite rest-frame $g - r$ color and absolute magnitude from the matched SDSS photometry. These filters leave 2,578 galaxies. Stellar-mass and gas measurements are not used.

For each completed catalogue, the bar label B_i is modeled by Bayesian logistic regression:

$$\operatorname{logit} \Pr(B_i = 1 \mid \mathbf{c}_i, \boldsymbol{\gamma}) = \gamma_0 + \gamma_M \widetilde{M}_{r,i} + \gamma_{M^2} \widetilde{M}_{r,i}^2 + \gamma_c \widetilde{c}_i + \gamma_z \widetilde{z}_i. \tag{11}$$

Here $\widetilde{M}_{r,i}$, $\widetilde{c}_i$, and $\widetilde{z}_i$ are standardized absolute magnitude, rest-frame $g - r$ color, and redshift. The quadratic magnitude term permits a non-monotonic relation, and redshift accounts for a possible observational trend. The intercept has prior $\mathcal{N}(0, 3^2)$, and the remaining coefficients have independent $\mathcal{N}(0, 1.5^2)$ priors.

The target comprises two conditional bar-frequency curves, giving 30 target coordinates. Each varies one of absolute magnitude or color over 15 points between its fifth and ninety-fifth percentiles, holding other predictors at their medians. Covariate standardization and the evaluation grids are fixed from the full analysis cohort. We form a Laplace approximation to the regression posterior for each completed catalogue and mix these posteriors. Gaussian quadrature integrates conditional curve means and variances. The oracle curves use the same integrated posterior mean fitted to all reference labels. Full review removes catalogue-label uncertainty, while regression uncertainty remains.

B.3.2 SPATIAL CLUSTERING

This task is inspired by (Skibba et al., 2009). The clustering task requires $0.01 < z < 0.08$ and $M_r < -20.5$, together with the morphology screening above and available positions and redshifts. It contains 9,050 galaxies and does not apply the face-on disk or bar-vote cuts. We convert angular coordinates and spectroscopic redshifts into comoving positions under flat ΛCDM with $\Omega_m = 0.3$. Distances are expressed in $h^{-1}\,\mathrm{Mpc}$, where $h = H_0/(100\,\mathrm{km\,s^{-1}\,Mpc^{-1}})$. Observed redshifts define redshift-space, not peculiar-velocity-corrected separations.

For class $c \in \{\mathrm{smooth}, \mathrm{featured}\}$, let $m_i^{(c)}$ be the binary class indicator, $\bar{m}_c$ its catalogue mean, and $\mathcal{P}_b$ the set of unordered galaxy pairs in separation bin b. Our target is

$$M_c(b) = \frac{\sum_{(i,\ell)\in\mathcal{P}_b} m_i^{(c)} m_\ell^{(c)}}{|\mathcal{P}_b|\,\bar{m}_c^2}. \tag{12}$$

We use three-dimensional separation bins with edges $0.1, 0.3, 1, 3, 10, 30\,h^{-1}\,\mathrm{Mpc}$ and concatenate both classes' statistics, giving ten target coordinates. A value above one indicates more same-class pairing than expected from the overall class fraction. This marked pair ratio conditions on the observed galaxy distribution and requires no survey random catalogue. It is distinct from a projected two-point correlation function. Empty pair bins or absent morphology classes yield undefined targets instead of artificial measurements.

We evaluate the nonlinear statistic on each completed morphology catalogue. Changing one label can affect many pairs and also changes the class-fraction normalization. Intervals quantify label uncertainty conditional on mean classifier probabilities, the selected galaxies, and their positions. They exclude classifier-parameter uncertainty, residual spatial label association, cosmic variance, and uncertainty in survey selection. The oracle is the same statistic computed with all reference labels, and label-induced uncertainty vanishes at full review.

B.3.3 MORPHOLOGY-ENVIRONMENT ASSOCIATION

We also estimate whether the frequency of galactic bars changes with local environment (Skibba et al., 2012). Our DECaLS implementation uses the same 2,578 face-on disk galaxies and bar labels

as the bar-frequency analysis. Local density is measured relative to a broader parent catalogue of 21,353 galaxies with finite positions, redshifts, and absolute magnitudes satisfying the same redshift and magnitude cuts. The parent catalogue is not restricted by morphology or embedding availability. With $r_{5,i}$ denoting the distance to the fifth other parent-catalogue galaxy, we set $d_i = 5/[(4\pi/3)r_{5,i}^3]$. Distances use the same redshift-space co-moving coordinates and flat ΛCDM cosmology as the clustering task.

For each completed catalogue, we fit a Gaussian-prior logistic regression and use its posterior-mode coefficients:

$$\operatorname{logit}\Pr(B_i = 1 \mid d_i, \mathbf{c}_i, \boldsymbol{\eta}) = \eta_0 + \eta_d \widetilde{\ell}_i + \eta_c \widetilde{c}_i + \eta_M \widetilde{M}_{r,i} + \eta_z \widetilde{z}_i, \tag{13}$$

where $\ell_i = \log_{10} d_i$, and tildes denote standardization over the set of samples. The adjustment covariates are rest-frame color, absolute magnitude, and redshift. This regression has no quadratic term. The intercept has prior $\mathcal{N}(0, 3^2)$, and the remaining coefficients have independent $\mathcal{N}(0, 1.5^2)$ priors.

The downstream target is a single adjusted change in mean bar probability between dense and sparse environments. Let $q_{0.25}$ and $q_{0.75}$ be the quartiles of ℓ_i in the analysis cohort. We set every galaxy's log density to either quartile while holding its observed color, magnitude, and redshift fixed. We estimate

$$T = \frac{1}{n}\sum_{i=1}^{n} \left[\Pr(B_i = 1 \mid \ell = q_{0.75}, \mathbf{c}_i, \widehat{\boldsymbol{\eta}}) - \Pr(B_i = 1 \mid \ell = q_{0.25}, \mathbf{c}_i, \widehat{\boldsymbol{\eta}})\right]. \tag{14}$$

Here $\widehat{\boldsymbol{\eta}}$ is refitted for each completed catalogue. This estimates a covariate-adjusted association, not a causal effect of environment. The oracle applies the same posterior-mode regression and contrast to all reference labels. Reported uncertainty propagates morphology-label uncertainty only, conditional on the observed covariates and density catalogue.

B.4 HUPD DATA AND TECHNOLOGY MEMBERSHIPS

We use raw HUPD application records, fixed to dataset revision `f570a84`, retaining first claims, complete CPC lists, filing dates, and inventor countries (Suzgun et al., 2023). We filter applications to have their filing date in 2015–2016 to have a nonempty first claim, usable CPC annotations, and complete inventor-country information. We normalize CPC codes to four-character subclasses and deduplicate memberships. We exclude missing classification lists. We set aside approximately 10% of applications for choosing the technology classes to study. Assignment depends only on the claim text, making the split reproducible.

Using this held-out subset, we choose 20 technology subclasses that PatentBERT can predict (see Appendix Figure 7). Each must have at least 100 applications in the subclass and 100 outside it. This separates class selection from the labels used to evaluate review strategies, while leaving most applications available for the experiments. The panel includes G06F, G06Q, H04L, A61K, and A61B, with the remaining classes chosen by frequency. From the remaining applications, we select 10,000 proportionally across filing years without class balancing, then reuse this catalogue across tasks, representations, and seeds. Country portfolios pool both years. The 15 countries with at least 100 retained applications receive fractional patent weights proportional to inventor counts. We retain the original fractions when other countries are excluded.

We obtain class membership PatentBERT logits from each first claim (Lee & Hsiang, 2020). It uses the pooled classification-token representation with a 128-token input limit. We apply label-free PCA to 32 dimensions, and use the shared scalar normalization. Independent Bayesian binary heads either learn class membership from these features or add a residual correction to PatentBERT logits. PatentBERT's reported training filings end in 2014, as such, our subset of 2015–2016 applications does not overlap.

B.4.1 PORTFOLIO DIVERSITY

Let Y_{ic} indicate whether patent i belongs to class c, and w_{ir} be the fraction of its inventors from country r. Country–class counts are $C_{rc} = \sum_i w_{ir} Y_{ic}$, with shares $p_{rc} = C_{rc} / \sum_d C_{rd}$. The target

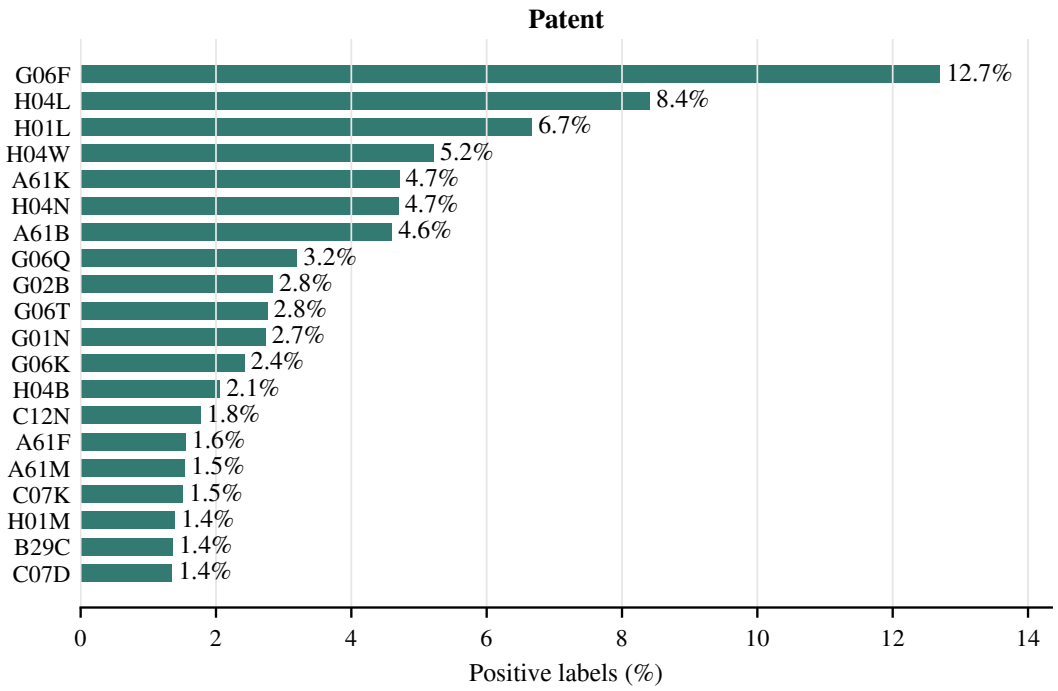

Figure 7: **Distribution of classes in the HUPD dataset.**

contains each country's normalized Shannon entropy,

$$H_r = -\frac{1}{\log K} \sum_{c=1}^{K} p_{rc} \log p_{rc}, \qquad K = 20. \tag{15}$$

Countries receive equal weight in evaluation, regardless of portfolio size. Each class membership contributes once, so patents with several classes contribute to several counts. We define entropy as zero for empty portfolios and record their posterior probability. Herfindahl concentration and effective class counts provide additional summaries. This target measures breadth within the selected class panel, not the diversity of all inventive activity.

B.4.2 Technology proximity

Let $m_c = \sum_i Y_{ic}$ and $m_{cd} = \sum_i Y_{ic} Y_{id}$ for a catalogue of N patents. Adding one half to each cell of the two-class contingency table gives the log-lift target

$$L_{cd} = \log \frac{(m_{cd} + 1/2)(N + 2)}{(m_c + 1)(m_d + 1)}, \qquad c < d. \tag{16}$$

The 190 coordinates measure co-occurrence relative to smoothed marginal frequencies. A review changes both the affected co-occurrences and their normalization. Conditional-probability proximity and country relatedness density are secondary summaries. We target edge strengths rather than community assignments, which can change discontinuously after small changes in the inferred network.

B.4.3 Comparative advantage

Let $p_{\cdot c} = \sum_r C_{rc} / \sum_{r,d} C_{rd}$ denote the pooled class share across included countries. Revealed technological advantage and the acquisition target are

$$\mathrm{RTA}_{rc} = \frac{p_{rc}}{p_{\cdot c}}, \qquad T_{rc} = \log(1 + \mathrm{RTA}_{rc})\,. \tag{17}$$

The target contains 300 country–class coordinates. Values of RTA above one indicate relative specialization, and their posterior probabilities are recorded separately. Empty portfolios and absent

classes have no revealed advantage and are flagged. This is comparative advantage within the retained patent catalogue, not a measure of productivity or a causal economic effect.

For all three tasks, nonlinear statistics are evaluated on sampled complete membership catalogues with reviewed entries fixed. Full review recovers the exact reference and removes label uncertainty. Unreviewed memberships are independent conditional on classifier parameters, while shared head uncertainty induces association across predictions. The model does not capture residual within-patent class association. All uncertainty statements condition on the retained applications, class panel, and inventor metadata.

C Acquisition approximations and review protocol

C.1 Acquisition approximations

We approximate the utilities in equation 1–equation 3 using current posterior uncertainty. Ecological EIG averages marginal mutual information across target coordinates, using discrete occupancy, richness, or count distributions. Conditioning on each possible answer retains direct local label association and effects across sites through shared coefficient draws. Galaxy acquisition uses linearized target sensitivities and analytic label moments, with a Gaussian covariance approximation for EIG in a reduced target representation. In both fixed and refit regimes, acquisition scores use the current classifier without hypothetical classifier retraining. Galaxy scores represent direct catalogue corrections, while ecological scores also condition on the current scientific model and calibration.

Patent acquisition samples four complete membership catalogues for each of 256 classifier draws. Each hypothetical answer forces only the candidate membership to zero or one, using common random numbers for all other entries. Sufficient-statistic updates propagate this change through country counts, co-occurrences, and their normalizations. EIG uses the shared Gaussian-mixture information approximation with a target projection of at most 30 dimensions. Branches retain the current classifier distribution, with no hypothetical classifier fits. Random, entropy, BALD, target variance, and Target EIG use the same reproducible uniform-shortlisting rule, evaluating at most 2,048 eligible pairs per round without consulting hidden labels. These are shortlist-based comparisons rather than exhaustive optimization over all 200,000 pairs.

C.2 Review rounds and classifier updates

Each experiment alternates inference, selection, and label revelation until its review budget is exhausted. We use a batch size of five for every task unless otherwise noted. Utility-based strategies rank candidates using one current posterior, while random review samples uniformly. Outside the split-label ablation (Appendix G.2), every review enters downstream scientific inference and refit classifiers use all acquired labels. Fixed ecological runs also freeze emission calibration, so that comparison changes both classifier and calibration updating.

C.3 Reference labels and oracles

Reviews are simulated by revealing existing annotations, while unrevealed answers and oracle estimates are excluded from fitting and acquisition. Oracle models use the same scientific model as partially reviewed models, just with all labels revealed.

D Model fitting with MCMC and variational inference

D.1 Choice of inference method

We fit Bayesian classifier heads using Markov chain Monte Carlo (MCMC) for the Galaxy Zoo DECaLS and HUPD tasks, and variational inference (VI) for Acoustic Forest Soundscape. The feature-extraction backbones remain frozen. Galaxy and patent scientific targets propagate label uncertainty through completed catalogues and task-specific population calculations, rather than fitting another model with MCMC. Spatial clustering conditions on mean classifier probabilities, while the other catalogue tasks also propagate head-parameter uncertainty. Following (Haucke et al., 2026b;

Rhinehart et al., 2022), acoustic tasks additionally fit a joint ecological model, separately from the classifier head. Both inference backends use NumPyro (Phan et al., 2019) with JAX (Bradbury et al., 2018) and double-precision arithmetic, with some utilities provided by Biolith (Haucke et al., 2026a).

Occupancy, richness, and abundance require repeatedly refitting an ecological model as expert reviews accumulate. Each fit uses the full survey dataset and jointly models ecological state, imperfect detection, and classifier score calibration across ten species. With five labels reviewed per round, acquiring 1,000 additional labels requires 200 posterior updates for each trajectory. VI makes these repeated ecological fits practical, and we validate its approximation against MCMC in Appendix D.3.

D.2 MCMC IMPLEMENTATION

We use NumPyro's No-U-Turn Sampler (NUTS), an adaptive Hamiltonian Monte Carlo algorithm (Hoffman et al., 2014). Each classifier fit runs three sequential chains, each with 500 warmup iterations for adaptation followed by 500 retained iterations, giving 1,500 posterior draws. The initial target acceptance probability is 0.9. Galaxy heads use a dense mass matrix, while the HUPD multilabel heads use a diagonal mass matrix. Each refit reruns warmup using all currently available labels. Fixed classifiers retain their initial posterior. Before any labels are available, pretrained residual heads use direct prior draws and do not require MCMC.

Classifier fits require zero divergent transitions, maximum split $\widehat{R} \leq 1.05$, and minimum effective sample size of 100 across sampled parameters, as computed by NumPyro. Fits that fail these checks are retried with target acceptance probabilities 0.97 and then 0.995, keeping the same chain and iteration counts.

D.3 VARIATIONAL IMPLEMENTATION FOR ACOUSTIC TASKS

We use full-rank Gaussian VI with automatic differentiation and reparameterized Monte Carlo gradients (Kucukelbir et al., 2017). Let $\boldsymbol{\xi}$ denote the continuous parameters of the fitted model, either classifier-head coefficients or ecological, detection, and score-calibration parameters. We optimize

$$q_\phi(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\xi} \mid \mathbf{m}, LL^\top), \tag{18}$$

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi}[\log p(\mathcal{D}_t, \boldsymbol{\xi}) - \log q_\phi(\boldsymbol{\xi})], \tag{19}$$

where L is a learned Cholesky factor. Within the ecological fit, the full covariance allows association between species, ecological coefficients, and score calibration. Positive emission separations and scales are parameterized on the log scale.

The likelihood marginalizes unreviewed detection labels and latent site states. For occupancy, we sum over the two possible presence states. For Royle–Nichols abundance, we integrate over latent counts using adaptive Poisson truncation (tolerance of 10^{-8}). VI and MCMC therefore target the same marginalized model. Predictions integrate the discrete site states conditional on each parameter draw and then average over draws.

We use NumPyro's `AutoMultivariateNormal` guide and `ClippedAdam`, with learning rate 0.01 and clipping threshold 10. Each model uses its full observed-data likelihood without minibatching. Occupancy and richness configurations use four Monte Carlo samples per gradient step, while abundance uses one. Each fit starts afresh with two independently perturbed initializations, and a separate 64-sample ELBO evaluation selects the better solution. Posterior samples are drawn from the selected guide.

We use 5,000 optimization steps, retain 1,000 posterior draws, and use 256 draws for prediction. The matched validation below compares these optimization lengths with longer fits. We retain optimization losses, restart ELBOs, posterior draws, and finite-value diagnostics. Finite optimization diagnostics do not establish posterior convergence. The validation conditions on saved classifier scores and evaluates only the ecological posterior approximation, not the separately fitted classifier heads.

Table 2: **The scientific posterior is sufficiently similar between 5,000-step VI and MCMC.** SD ratios are VI divided by MCMC, with one median across target coordinates.

Target	Observations reviewed	Posterior-mean MAE	Median SD ratio
Occupancy	Partial	0.0018–0.0036	1.001–1.026
Richness	Partial	0.0073–0.0279	1.002–1.008
Abundance	Partial	0.0182–0.0278	1.001–1.010
Occupancy	Oracle	0.0025	1.002
Richness	Oracle	0.0169	1.002
Abundance	Oracle	0.0220	0.996

Table 3: **VI is substantially faster on abundance inference.** Entries are in seconds. VI includes two restarts and MCMC includes four chains. All columns exclude acquisition scoring and classifier training.

Regime	VI 5k	VI 10k	VI 20k	MCMC
0 reviews	90	107	146	2,470
Oracle	79	90	118	1,173

D.4 Matched MCMC and VI validation

We validate VI on Acoustic Forest Soundscape using the 104 sites, 1,302 recordings, and ten-species panel described in Appendix B.1. For both occupancy and Royle–Nichols abundance, we evaluate multiple Target-EIG trajectories. Within each trajectory, VI and MCMC use the same classifier scores and normalization, reviewed-label mask, covariates, species panel, likelihood, and priors. Oracles condition on all observation labels and omit score-emission calibration. We compare 5,000-, 10,000-, and 20,000-step fits, each using two restarts and 3,000 posterior draws. For prediction and acquisition calculations, all methods use 512 posterior draws.

For this validation, the NUTS reference uses a dense mass matrix and increases sampling to four sequential chains, 1,000 warmup iterations and 1,500 retained iterations per chain, with target acceptance probability 0.95. This gives 6,000 retained draws before subsampling for prediction and acquisition. We require zero divergences, maximum rank-normalized split and folded $\widehat{R} \leq 1.01$, and minimum bulk and tail effective sample sizes of 400 (Vehtari et al., 2021). All fits pass these checks on the first attempt. The largest $\widehat{R}$ is 1.0031, the smallest bulk effective sample size is 4,963, and the smallest tail effective sample size is 3,702.

D.5 Posterior agreement and computational cost

As shown in Appendix Table 2, VI closely matches MCMC for occupancy, richness, and abundance, both during active review and after full review. Posterior uncertainty for the scientific targets is also well preserved. Median target standard deviations agree within 2.6% for occupancy and within 1% for richness and abundance. The oracles show similarly close agreement.

Agreement is somewhat weaker for individual abundance and detection coefficients, where VI modestly underestimates posterior uncertainty. However, the two methods generally lead to the same qualitative conclusions about these coefficients and preserve the association between abundance and detection. Thus, the approximation is strongest for the downstream quantities used in our experiments, even when some parameter-level differences remain.

VI is substantially faster. As shown in Appendix Table 3, a 5,000-step VI fit takes roughly one to two minutes, compared with tens of minutes for the matched MCMC reference. This reduction makes repeated posterior updates practical during active review.

D.6 Does VI preserve acquisition utility?

Close posterior means do not guarantee that an uncertainty-sensitive review policy will make similar choices. We therefore score the same complete pool of eligible species–recording labels under each

Table 4: **At 5,000 steps, VI retains MCMC-estimated acquisition utility.** Entries are percentages from Eq. (20).

Target	VI 5k	VI 10k	VI 20k
Occupancy	100.0	99.1	100.0
Richness	100.0	100.0	100.0
Abundance	98.4	100.0	99.1

posterior. Let $\bar{u}_{\mathrm{MC}}(i)$ denote the ecological EIG score under MCMC, averaged over scientific target coordinates as described in Appendix C. Let B_q and B_{MC} contain the five candidates with the largest individual utilities under VI and MCMC, respectively. We measure the retained utility

$$\mathcal{R}(B_q) = \frac{\sum_{i \in B_q} \bar{u}_{\mathrm{MC}}(i)}{\sum_{i \in B_{\mathrm{MC}}} \bar{u}_{\mathrm{MC}}(i)}. \tag{20}$$

Both selected batches are evaluated under the same MCMC posterior. This tests the sum of individual marginal utilities implemented by the review loop. It does not estimate joint batch information or the realized scientific improvement after review and classifier refitting. We also compare Spearman correlations of all eligible scores and overlap between the selected batches. A retained utility of at least 95% is a practical tolerance specified for this validation, not a universal posterior-accuracy criterion.

At 5,000 steps, VI preserves nearly all of the acquisition utility obtained under MCMC. Candidate rankings are highly similar, even when the exact five selected examples differ. Increasing the number of VI optimization steps does not consistently improve these decisions. We therefore use 5,000-step VI in the main experiments.

E Additional baselines

In order to make fair comparisons, we compare Target EIG against baseline strategies that use the same inference model and differ only by which examples they prioritize. Active statistical inference (Zrnic & Candes, 2024) and active measurement (Hamilton et al., 2026) instead combine acquisition with their own unbiased estimators, which apply to means, totals, and regression coefficients but not to latent quantities like site occupancy or to targets built from joint labels such as technology proximity. Active measurement additionally relies on a sampling distribution shaped to each observation's contribution to a population total, which has no natural analogue when the target is a latent state or depends jointly on several labels. We therefore provide a limited comparison to Active Statistical Inference as the closest conceptual baseline for tasks it supports.

E.1 Active statistical inference

Target EIG selects reviews to reduce posterior uncertainty in a scientific target. Active statistical inference (ASI) (Zrnic & Candes, 2024) combines selective labeling with prediction-error correction. We implement it for DECaLS bar frequency and environmental association, and HUPD portfolio diversity and comparative advantage. Galaxy estimates use corrected logistic regression with the same Gaussian-prior penalty as the fully labeled reference, followed by fitted bar probabilities or their adjusted environment contrast. For HUPD, corrected fractional country–class totals are transformed into normalized entropy or $\log(1 + \mathrm{RTA})$. Technology proximity is excluded from this implementation because its co-occurrence totals involve joint memberships, whereas one review reveals only one membership.

ASI scores multiply classifier uncertainty by the root mean squared influence on the scientific target, using regression derivatives for galaxies and count-transformation derivatives for patents. All comparisons keep the pretrained classifier fixed. Before sampling, we freeze predictions and Bernoulli inclusion probabilities, mixing targeted and uniform allocation equally. Each requested budget specifies an expected label cost, so the realized number of reviews varies. Reviewed labels contribute inverse-probability-weighted prediction-error corrections. A uniform-sampling control uses the same corrected estimator.

For HUPD, these corrections can yield negative totals or totals larger than the available country weight. We therefore clip each aggregate country–class count to $[L_{rc}, U_{rc}]$, where L_{rc} is the confirmed positive weight and U_{rc} adds the weight of all unreviewed memberships for that country and class. The bounds use only acquired labels, including initialization labels, and observed inventor-country weights. They preserve nonnegative counts before the entropy and RTA transformations, without clipping individual prediction-error corrections. It should be noted that this projection sacrifices unbiasedness. Delta-method intervals are withheld for affected or nonregular targets, except for a probability-one census. Galaxy intervals instead use a regression sandwich covariance and the delta method. We compare point errors here, without claiming coverage of these asymptotic, catalogue-conditional intervals.

Neither approach dominates uniformly (Appendix Figure 8). For galaxy bars near 500 labels, ASI achieves RMSE 0.0328, compared with 0.0450 for Target EIG and 0.0507 for Bayesian random review. For HUPD portfolio diversity near 2,000 labels, bounded ASI has RMSE 0.0831, compared with 0.0336 for Target EIG and 0.1035 for uniform corrected inference. ASI can improve on uniform corrected inference, but is variable at small budgets. These comparisons change both acquisition and estimation, so they do not isolate acquisition quality.

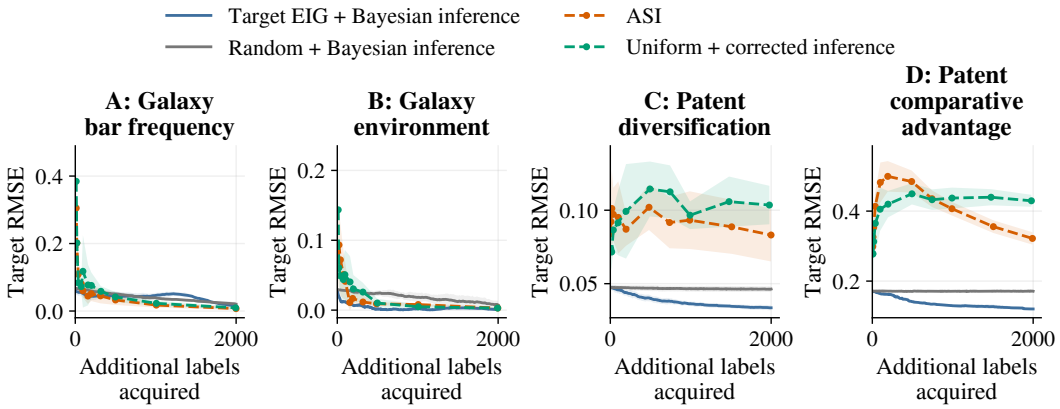

Figure 8: **Target EIG and active statistical inference offer different tradeoffs.** All panels use fixed pretrained classifiers. Lines and shading show means and sample standard deviations over five complete seeds. ASI points summarize separate completed streams for each requested budget at their mean actual label cost. Bayesian curves are nested trajectories. Galaxy errors use a common plug-in regression target. HUPD corrected-inference results use bounded country–class counts. Errors are RMSE in each target's native units, on linear axes. Curves without five complete finite seeds are omitted.

F Runtime Analysis

We compare observed runtimes for the downstream model fitting and target-directed sampling used in Figure 3, across tasks. We average recorded rounds after initialization and before 200 reviews within each run, then average matched seeds. The first round and final evaluation are excluded.

Appendix Figure 10 measures only the wall time required to score and select the next review batch. It excludes fitting, prediction, and evaluation stages. Random and classifier-directed selection are generally inexpensive, while Target EIG incurs additional task-dependent scientific calculations. These timings do not include the preceding prediction stage that produces classifier probabilities. Appendix Figure 9 shows the time scientific model fitting takes, averaging strategies equally within each seed.

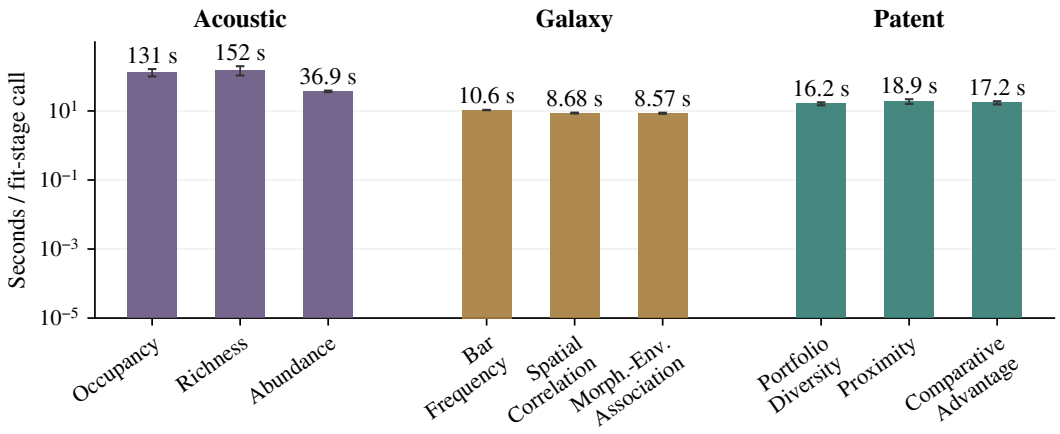

Figure 9: **Model fitting runtimes vary across tasks.** Bars show the mean time spent fitting per round, with strategies weighted equally within each seed. Colors identify datasets and error bars show sample standard deviations across seeds. Prediction and scientific calculations are excluded. The vertical axis is logarithmic.

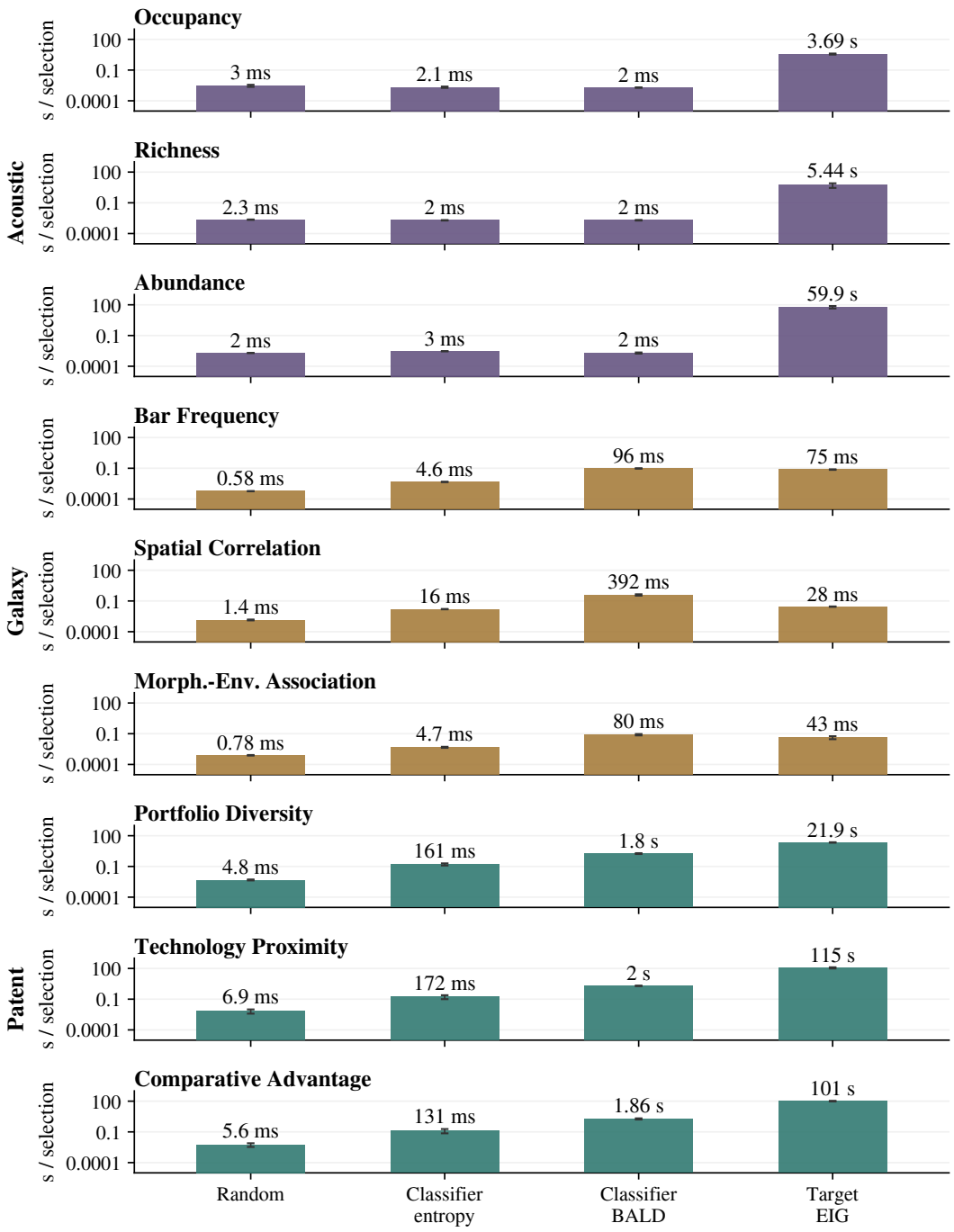

Figure 10: **Review-strategy runtime depends strongly on the scientific task.** Each row reports mean recorded wall time per review round, including fitting, prediction, selection, and evaluation. Colors identify datasets. Error bars show sample standard deviations across matched seeds and are absent for single-seed acoustic runs. Axes are logarithmic, and annotations give absolute times. Batch sizes and seed counts are shown for each task.

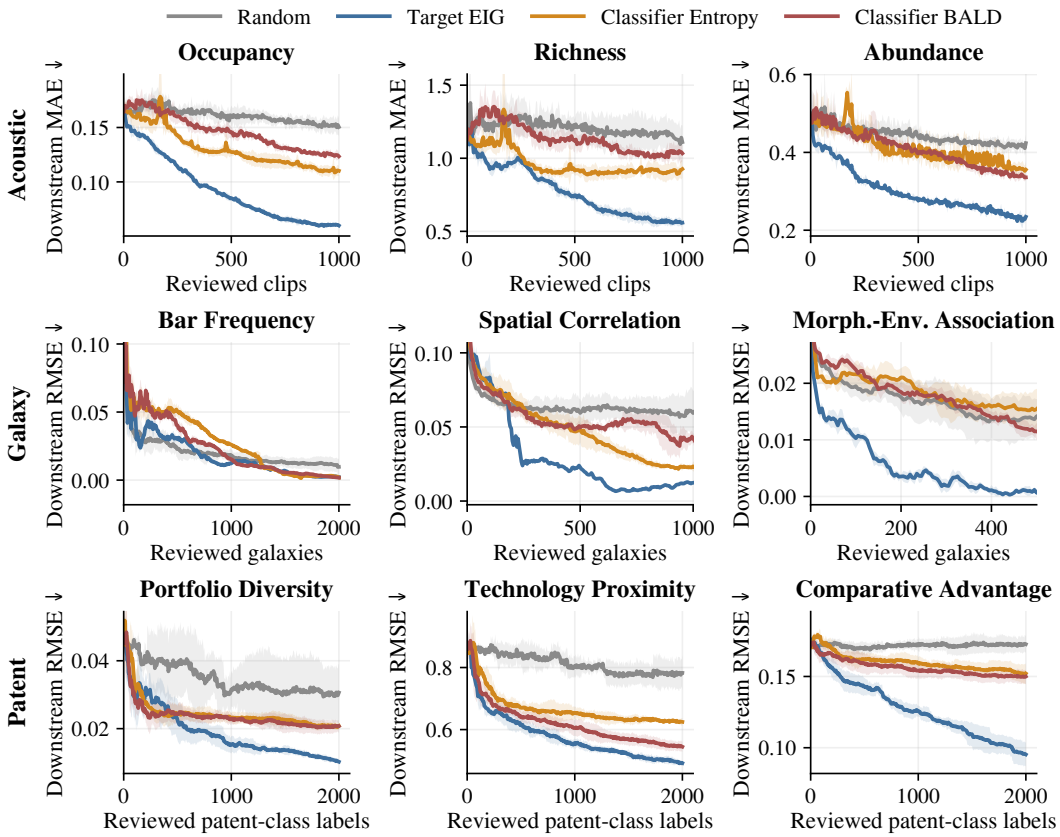

Figure 11: **Even if refitting the classifier, prioritizing samples for the downstream scientific inference using Target EIG yields lower error than classifier-centric review for eight out of nine tasks.** Like for Figure 3, bar frequency is the outlier (see Section 4.2). Error relative to the oracle versus number of expert-reviewed labels; lines show means and shading ± 1 SD across seeds. Error scales differ across tasks due to variable dataset and target difficulty.

G Ablations and additional results

G.1 Refit review trajectories

G.2 Splitting reviews between classifier and scientific updates

The mixed schedules allocate half the task-specific review budget to classifier BALD and half to Target EIG, either sequentially or in alternating batches of five. BALD labels are used only for classifier fitting, and EIG labels only for direct scientific-model updates. Updated classifier predictions can still indirectly change downstream scientific inference. The all-EIG baseline instead uses every acquired label for both roles. These comparisons therefore vary both acquisition and for which purpose labels are used.

G.3 When does classifier refitting improve galaxy inference?

Classifier refitting and review selection can improve inference in two different ways. Reviewed labels help *directly* by replacing uncertain morphology predictions in the downstream analysis. Refitting can add an *indirect* gain by improving predictions for galaxies that were not reviewed.

We study the effect of these on the galaxy dataset, comparing the original classifier predictions, the same data with *reviewed* labels, and the after additionally refitting the classifier. We then measure how much downstream error is reduced directly by review, how much is reduced through refitting,

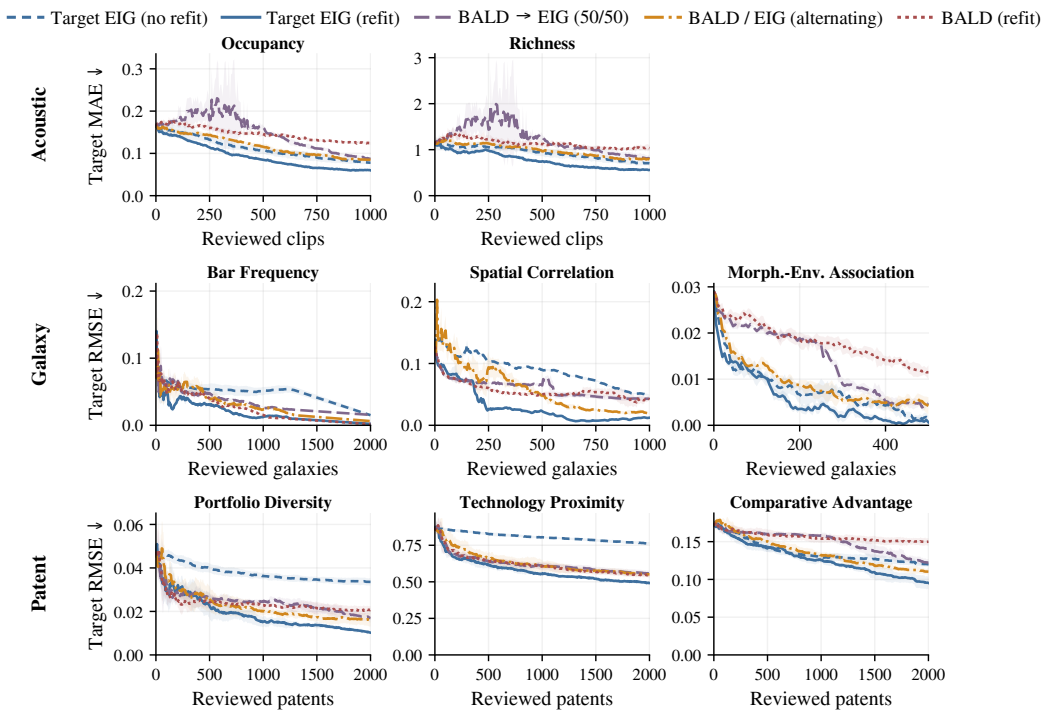

Figure 12: **Refitting helps most when using labels selected for the scientific target.** Downstream error for all-EIG review, BALD followed by EIG, and alternating batches, and all-BALD review as a function of expert-reviewed labels. Lines and shading show means and ±1SD across seeds.

which parts of the task-relevant data distribution are covered by the reviewed galaxies, and whether refitting moves the target toward the oracle.

When can classifier-directed refitting overcome weaker review selection? BALD can outperform fixed Target EIG when improvements to the classifier propagate broadly enough to compensate for reviewing less scientifically informative galaxies. Spatial clustering is the clearest example: Target EIG directly covers much more of the target-sensitive data, but BALD refitting substantially improves predictions across the remaining catalogue. Because each morphology prediction contributes to many galaxy pairs, this propagated improvement can dominate the direct value of the reviewed labels.

The morphology–environment task shows the opposite case. Target EIG reviews galaxies that directly improve the downstream estimate, and BALD's broader classifier improvement is not large enough to make up for missing those observations. Thus, classifier-directed refitting is useful only when its propagated benefit exceeds the downstream value sacrificed by not reviewing the most target-relevant examples.

Why does refitting Target EIG help? Target-EIG reviews can provide both kinds of value: they directly constrain the scientific target and also provide supervision that improves predictions for unreviewed galaxies. Refitting therefore helps most when those prediction changes generalize broadly and move the downstream estimate toward the oracle.

This effect is not simply that refitting improves classification more on "task-relevant" galaxies. We see only modest evidence that predictive improvements are concentrated on highly target-sensitive examples. The stronger signal is alignment in target space: across all three galaxy tasks, Target-EIG refitting moves the downstream estimate toward the oracle. What matters for downstream scientific inference is therefore not only whether the classifier improves, but whether its improvements occur in ways that matter for the target.

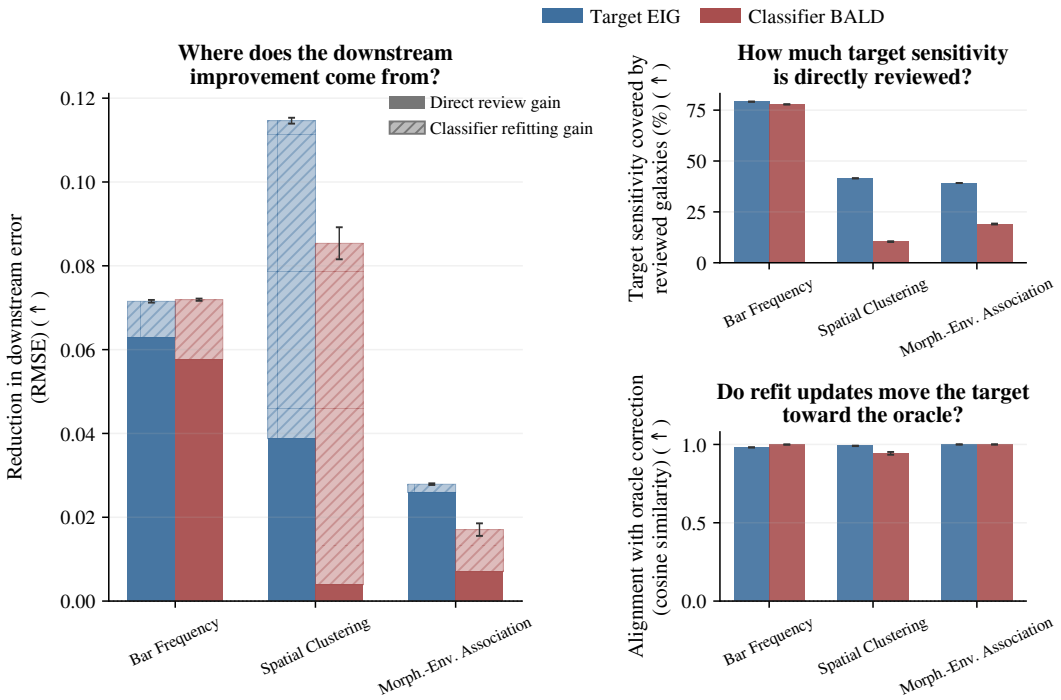

Figure 13: **Refitting helps when improvements to unreviewed predictions move the scientific target in the right direction.** Left: downstream RMSE reduction after the full review trajectory, split into the direct gain from inserting reviewed labels and the additional gain from refitting the classifier. Top Right: fraction of total target sensitivity covered by the reviewed galaxies. Bottom Right: cosine similarity between the target change caused by refitting and the correction needed to reach the oracle, where one indicates perfect alignment. Bars show means over five seeds and error bars show one standard deviation.

G.4 Review batch size

We vary the number of labels acquired before refitting, $b \in \{1, 5, 25\}$, using the pretrained classifier with Target EIG. We compare error at task-specific endpoints: 1,000 reviews for acoustic tasks and spatial correlation, and 2,000 for bar frequency and patent tasks. Smaller batches adapt more frequently to newly reviewed labels, at the cost of more frequent fitting.

Available acoustic and bar-frequency comparisons show modest differences across batch sizes (Appendix Table 5). Patent portfolio diversity benefits more from frequent updates, with mean error increasing from 0.0099 at batch size one to 0.0167 at 25. Unless otherwise noted, we use batches of five for all experiments as a compromise between runtime and accuracy.

G.5 Extended Analysis of Selected Data

Appendix Table 6 shows that across the acoustic tasks, Target EIG prioritizes positive detections over uncertain negatives. This is even more the case for the abundance task, where the goal is not just to establish species presence, but the number of presence observations. For bar-fraction regression, it favors galaxies at covariate extremes, where labels have greater leverage on the fitted linear relationships. For spatial correlation, it instead selects galaxies near previously reviewed examples, creating informative labeled pairs. Thus, even within the same dataset and classification problem, the *downstream target* determines which observations are most valuable to review.

For the patent tasks, the target definitions imply different sources of label value. Membership in a small country portfolio can substantially change normalized entropy. The same membership can affect several proximity edges through the patent's other classes, while also changing marginal

Task	n	$b = 1$	$b = 5$	$b = 25$
Acoustic occupancy	5	0.058 ± 0.001	0.059 ± 0.001	0.058 ± 0.003
Acoustic richness	5	0.560 ± 0.019	0.548 ± 0.027	0.537 ± 0.032
Acoustic abundance	1	0.254	0.238	0.254
DECaLS bar frequency	5	0.002 ± 0.000	0.002 ± 0.000	0.002 ± 0.000
DECaLS spatial correlation	5	0.011 ± 0.001	0.012 ± 0.001	0.010 ± 0.001
Patent diversification	5	0.010 ± 0.001	0.012 ± 0.002	0.017 ± 0.003
Patent proximity	5	—	0.498 ± 0.005	0.561 ± 0.023
Patent comparative advantage	5	—	0.092 ± 0.007	0.119 ± 0.011

Table 5: **The effect of review batch size varies across tasks.** Error at each task's default budget for batches of 1, 5, or 25 Target-EIG labels, with pretrained classifiers and refitting. Values are means $\pm$ sample standard deviations across matched seeds (lower is better). Seed counts are given as n. Single-seed estimates have no standard deviation.

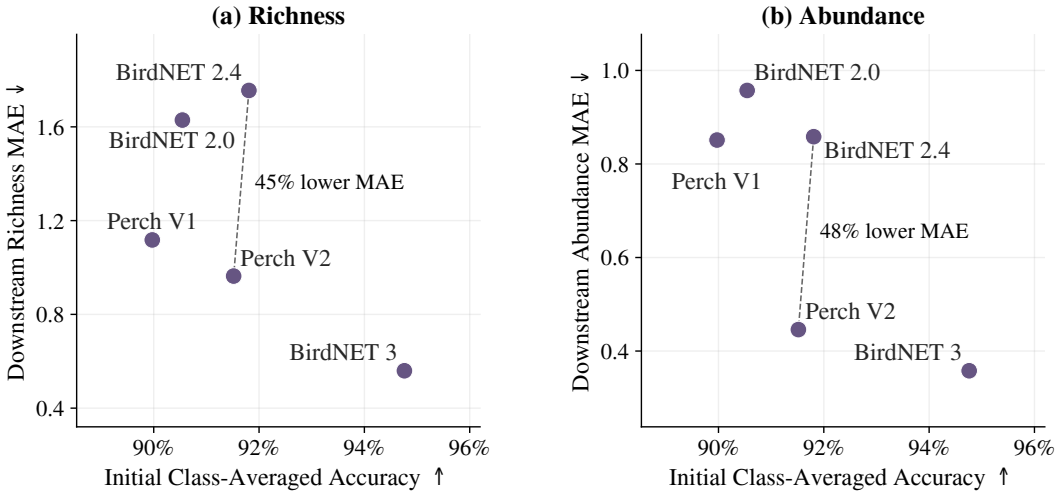

Figure 14: **Classifier accuracy and downstream scientific error are not perfectly correlated.** This figure is showing the relationship between classifier accuracy and downstream scientific inference error on the richness and abundance tasks. See Figure 2 for the corresponding occupancy results and Figure 4 for the corresponding review trajectories.

Dataset	Properties of Selected Data	Downstream Inference Target	Random	Classifier entropy	Target EIG
Acoustic	Distinct sites reviewed	Occupancy	101.00 ± 0.88	98.80 ± 0.56	**104.00** ± 0.00
		Richness	98.60 ± 3.12	98.80 ± 0.56	**104.00** ± 0.00
		Abundance	101.00	98.50 ± 6.35	**104.00**
	Positive labels (%)	Occupancy	15.16 ± 1.41	1.20 ± 0.28	**18.14** ± 0.68
		Richness	16.64 ± 0.80	1.20 ± 0.28	**18.16** ± 0.64
		Abundance	16.80	1.20 ± 3.81	**29.50**
HUPD	Previously reviewed patents (%)	Portfolio Diversity	9.00 ± 0.21	16.05 ± 0.38	51.66 ± 0.77
		Technology Proximity	9.00 ± 0.21	16.05 ± 0.38	**50.81** ± 0.91
		Comparative Advantage	9.00 ± 0.21	16.05 ± 0.38	39.46 ± 1.27
	Embedding distance to prior reviewed patents	Portfolio Diversity	0.123 ± 0.001	0.096 ± 0.001	**0.141** ± 0.002
		Technology Proximity	0.123 ± 0.001	0.096 ± 0.001	0.100 ± 0.002
		Comparative Advantage	0.123 ± 0.001	0.096 ± 0.001	**0.140** ± 0.002

Table 6: **Target-directed and classifier-directed acquisition prioritize different examples.** Rows with selection characteristics deemed beneficial for the task are colored (see Appendix B for detailed definitions and justifications). Bold values highlight the best selection characteristic in each row.

frequencies. For comparative advantage, it changes both a country's class share and the pooled reference share.