BEYOND A SINGLE SPARSE SOLUTION: CERTIFYING FEATURE NECESSITY UNDER INEXACT MULTI-LABEL SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feature selection under inexact multi-label supervision is complicated by both ambiguous label assignments and redundant predictive supports. Existing methods typically identify the support of a single fitted solution, which may be predictive but does not establish whether its selected features are necessary. We introduce FIND—**F**eature necessity under **IN**exact supervision via **D**eletion—to study feature necessity over the family of near-optimal predictors. Our framework evaluates a feature set through counterfactual deletion: the set is removed, all remaining variables are re-optimized, and the resulting deletion cost determines whether any comparably good predictor can avoid it. This formulation characterizes both individual necessity, where a feature must appear in every near-optimal predictor, and collective non-removability, where individually replaceable features cannot be removed jointly. We instantiate the framework with a convex ambiguity-aware objective that preserves set-valued supervision and induces row-sparse feature selection. Strong convexity yields safe bounds on optimization error and deletion costs, enabling safe screening and selective re-optimization, while monotonicity supports structured search for collective relevance. Extensive experiments on controlled and real-world datasets validate the predicted individual and collective structures and demonstrate the practical effectiveness of the resulting certificates and feature rankings.

1 INTRODUCTION

Multi-label learning (MLL) is widely used in image annotation, music annotation, and disease prediction, where an instance may convey several semantic concepts (Duygulu et al., 2002; Turnbull et al., 2008; Pham et al., 2022). Precise multi-label annotation is both time-consuming and expensive because each instance may be associated with an unknown number of relevant labels. Weakly supervised MLL has therefore become a major paradigm for addressing this challenge. Two representative paradigms are partial multi-label learning (PML), whose candidate set contains all relevant labels and possibly irrelevant ones (Xie & Huang, 2018), and complementary multi-label learning (CML), whose annotations identify irrelevant labels. Since non-candidate labels in PML are complementary labels, and vice versa, the two paradigms can be unified as MLL under inexact supervision (Wang et al., 2026). Another issue of multi-label data is that it usually combines high-dimensional feature spaces with sparse discriminative information. Feature selection is therefore important for effective, efficient, and interpretable learning under such supervision.

Existing PML feature selection methods address label ambiguity from different perspectives. They reconstruct credible labels, exploit latent feature–label structures, or jointly learn label confidences and a row-sparse projection (Zhang & Fang, 2020; Hao et al., 2023; Pan et al., 2025; Gao et al., 2025). Despite these differences, they commonly produce a label–feature explanation $(\widehat{Y}, \widehat{W})$ that fits the observed partial labels, and infer feature importance from the coefficient matrix. Since the true labels are unobserved, satisfying the candidate-label constraints identifies a feasible label assignment rather than validating it. Correlations and redundancies among features allow different feature subsets to support the same label assignment. The observed data may therefore accommodate multiple label–feature explanations. When label confidences and feature weights are determined together, these two degrees of freedom can reinforce each other: a feasible label explanation favors

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.

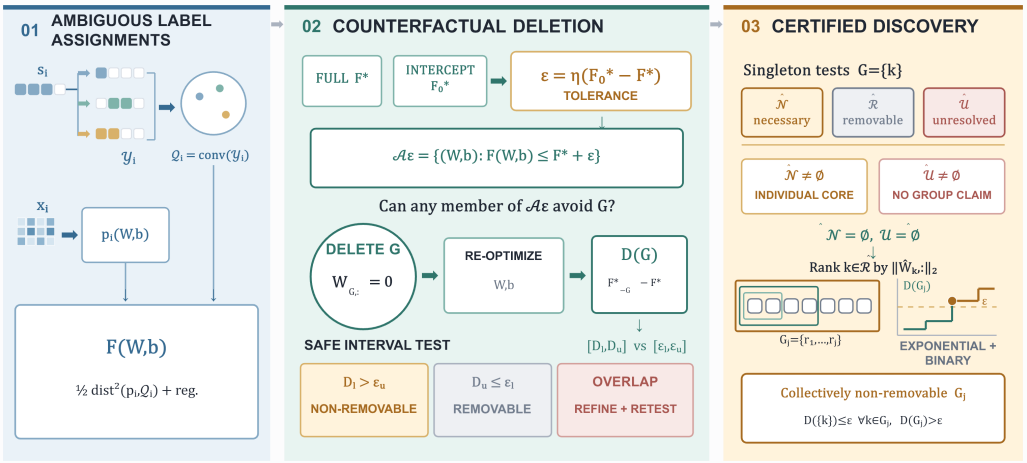


Figure 1: Overview of the proposed ambiguity-aware set-deletion framework. Set-valued annotations preserve admissible label assignments; counterfactual deletion then certifies whether an individual feature or a feature set can be avoided within the near-optimal family.

the features that best reproduce it, while the predictions made by those features, in turn, support the same label explanation. The selected support is therefore best viewed as the feature subset associated with one admissible explanation, even when it achieves strong predictive performance.

Such predictive performance is valuable when the sole objective is to construct an accurate predictor. It nevertheless establishes only that there exists a successful model based on the selected feature subset. Feature necessity asks a stronger question: whether a selected feature must remain active when all other models that fit the same observed supervision nearly as well are also considered. Specifically, a feature is necessary if excluding it prevents any admissible model from attaining a near-optimal value of the learning objective. Necessary features therefore capture the part of feature relevance that persists across all near-optimal explanations, rather than the support chosen by one successful predictor. Distinguishing these two notions allows feature selection to provide not only a useful subset for prediction, but also stronger evidence about which features are consistently supported by the observed supervision. This leads to our central question: beyond finding one model that predicts well, can we determine which features every near-optimal predictor must retain?

To answer this question, we evaluate feature relevance through counterfactual deletion. For a feature set G, we compare the optimum with all features against that after G is disabled and all remaining parameters are re-optimized. If the deterioration exceeds a prescribed near-optimal tolerance, no predictor of comparable objective quality can avoid G. A singleton G therefore identifies an individually necessary feature, whereas a larger G identifies a collectively non-removable set whose members may substitute for one another. This view shifts feature importance from the coefficient pattern of one fitted model to the best alternative explanation after deletion.

Figure 1 summarizes this pipeline. We refer to this framework as FIND. We instantiate the deletion principle with a convex ambiguity-aware feature selector. A squared-distance loss retains each annotation as a convex set of admissible targets instead of fixing pseudo-labels, while an $\ell_{2,1}$-regularized linear predictor selects rows shared across labels. Strong convexity yields safe optimality and deletion-cost bounds. We calibrate the tolerance by the feature-attributable improvement over an intercept-only model, screen deletion with certified bounds, and re-optimize only unresolved cases. The procedure returns an individual necessity core, a predictive ranking, and, when singleton features are replaceable, a nested search for collectively non-removable sets. Our contributions include:

- Formulating feature necessity under set-valued multi-label supervision over a near-optimal model family, distinguishing relevance shared across sufficiently good predictors from the support of one fitted solution.

- Developing a set-deletion framework for characterizing individual necessity and collective non-removability, with a convex ambiguity-aware instantiation and safe deletion-cost certification.
- Empirically validating the resulting individual and collective relevance structures on controlled and real-world PML datasets, while demonstrating competitive predictive performance under standard feature-budget protocols.

2 Problem Setup and Ambiguity-Aware Objective

Let $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n\times d}$ denote the feature matrix, where each instance has d features, and let $S = [s_1, \ldots, s_n]^\top \in \{0,1\}^{n\times q}$ denote the observed inexact annotations over q labels. In PML, $s_{ij} = 1$ indicates that label j is a candidate of instance i, and every relevant label is assumed included in the candidate set. A CML annotation can be written in the same form by setting $s_i = \mathbf{1} - c_i$, where c_i marks the observed complementary labels. The true label vector is unavailable during training. We assume that every instance has at least one relevant label and a nonempty candidate set.

For a nonempty candidate set s_i, the binary label assignments consistent with the observation are

$$\mathcal{Y}_i = \left\{ y \in \{0,1\}^q : 0 \le y \le s_i,\ \mathbf{1}^\top y \ge 1 \right\}. \tag{1}$$

Instead of selecting one element of $\mathcal{Y}_i$ as a pseudo-label, we retain its convex hull

$$\mathcal{Q}_i = \operatorname{conv}(\mathcal{Y}_i) = \left\{ z \in [0,1]^q : 0 \le z \le s_i,\ \mathbf{1}^\top z \ge 1 \right\}. \tag{2}$$

Thus, $\mathcal{Q}_i$ contains all convex label assignments compatible with the observed annotation.

We consider a linear multi-label predictor $p_i(W,b) = W^\top x_i + b$, where $W \in \mathbb{R}^{d\times q}$ and $b \in \mathbb{R}^q$. Its ambiguity-aware loss is defined by the squared distance to the feasible target set:

$$\ell_i(W,b) = \frac{1}{2}\operatorname{dist}^2\big(p_i(W,b), \mathcal{Q}_i\big) = \min_{z_i \in \mathcal{Q}_i} \frac{1}{2}\|W^\top x_i + b - z_i\|_2^2. \tag{3}$$

The loss does not commit to a fixed reconstruction of the unknown labels. Instead, each prediction is compared with the closest target compatible with the observed candidate set. Since this compatibility alone cannot identify the true labels, the shared predictor and its regularizers determine which explanations are preferred. Accordingly, our certificates concern feature necessity under this learning objective, not recovery of the unknown labels.

The feature-selection objective is

$$F(W,b) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(W,b) + \lambda\|W\|_{2,1} + \frac{\gamma}{2}\left(\|W\|_F^2 + \|b\|_2^2\right), \tag{4}$$

where $\|W\|_{2,1} = \sum_{k=1}^{d} \|W_{k,:}\|_2$ promotes a feature support shared across labels (Nie et al., 2010; Yuan & Lin, 2006). The ridge term with $\gamma > 0$ is introduced for numerical certification: it makes F strongly convex and later enables computable bounds on optimization error and deletion costs.

This objective supplies a common criterion for admissible predictive quality. Rather than reading feature relevance from its optimizer, we next use it to test whether a feature set can be removed while some predictor remains near-optimal.

3 Certifying Feature Necessity under Inexact Supervision

3.1 Near-Optimal Avoidability and Deletion Costs

For a feature set $G \subseteq \{1, \ldots, d\}$, we evaluate necessity by whether constraining $W_{G,:} = 0$ still permits a predictor of comparable objective quality. Let

$$F^\star = \min_{W,b} F(W,b) \tag{5}$$

and define the near-optimal model family at tolerance $\epsilon \ge 0$ as

$$\mathcal{A}_\epsilon = \{(W,b) : F(W,b) \le F^\star + \epsilon\}. \tag{6}$$

Here, $F^\star$ serves only as a reference level for comparable objective quality; feature relevance is not inferred from the optimizer attaining it. Instead, we determine whether the constraint excluding G remains compatible with the near-optimal family:

$$\exists\,(W,b) \in \mathcal{A}_\epsilon \quad \text{such that} \quad W_{G,:} = 0\,? \tag{7}$$

If such a model exists, every feature in G can be avoided simultaneously without leaving the near-optimal family. Otherwise, every near-optimal predictor must retain at least one member of G. Accordingly, feature necessity is determined by the availability of near-optimal alternatives after deletion, rather than by the support of a particular fitted model.

Whether G is avoidable therefore depends on how much objective deterioration is permitted when defining a comparable alternative. The tolerance ϵ makes this allowance explicit: if $0 \le \epsilon_1 \le \epsilon_2$, then $\mathcal{A}_{\epsilon_1} \subseteq \mathcal{A}_{\epsilon_2}$. Increasing ϵ enlarges the family in which a predictor avoiding G may be found, and hence makes G easier to declare removable. Conversely, if no such predictor exists even at a larger tolerance, the resulting non-removability claim is stronger.

Directly searching the entire family $\mathcal{A}_\epsilon$ for a model satisfying $W_{G,:} = 0$ is impractical. We instead solve the counterfactual optimization problem

$$F^\star_{-G} = \min_{W_{G,:}=0,\, b} F(W,b) \tag{8}$$

and define the set-deletion cost

$$D(G) = F^\star_{-G} - F^\star. \tag{9}$$

All parameters outside G are re-optimized in $F^\star_{-G}$. This step is essential: simply zeroing $W_{G,:}$ in one fitted model measures the damage to that model and can overstate the importance of G, because correlated features are not allowed to adapt. In contrast, $D(G)$ measures the smallest deterioration achievable by any predictor that avoids G.

The deletion cost provides an exact test for equation 7.

Theorem 1 (Near-optimal avoidance)**.** *For any $G \subseteq \{1,\dots,d\}$ and $\epsilon \ge 0$, $D(G) \le \epsilon$ if and only if there exists $(W,b) \in \mathcal{A}_\epsilon$ satisfying $W_{G,:} = 0$. Equivalently,*

$$D(G) > \epsilon \quad \Longleftrightarrow \quad \|W_{G,:}\|_F > 0 \quad \textit{for every } (W,b) \in \mathcal{A}_\epsilon. \tag{10}$$

Thus, $D(G)$ is not merely another importance score; it is a computable answer to whether any near-optimal predictor can avoid G. This characterization requires only that the relevant optima are attained. Strong convexity is not part of the semantic definition of necessity and is used only for efficient numerical certification in Section 3.2. The proof is given in Appendix D.3.

The meaning of ϵ depends on the scale of the learning objective. We calibrate it against the total objective improvement attributable to the feature space. Let

$$F^\star_0 = \min_b F(0,b) \tag{11}$$

be the intercept-only optimum and define the feature-attributable objective gain

$$B = F^\star_0 - F^\star. \tag{12}$$

For $B > 0$, we set

$$\epsilon = \eta B, \qquad 0 < \eta \le 1, \tag{13}$$

and define the normalized deletion cost

$$\rho(G) = \frac{D(G)}{B}. \tag{14}$$

A set is non-removable when deleting it sacrifices more than an η-fraction of the feature-attributable objective gain. Since the intercept-only model remains feasible after any partial feature deletion,

$$D(\varnothing) = 0, \qquad D(\{1,\dots,d\}) = B, \qquad 0 \le D(G) \le B. \tag{15}$$

Consequently, $\rho(G) \in [0,1]$ and $D(G) > \epsilon$ is equivalent to $\rho(G) > \eta$. Unlike λ and γ, η does not change the baseline predictor; it controls only how much deterioration is allowed when determining whether an alternative predictor remains near-optimal. If $B = 0$, the features offer no improvement over the intercept-only model and no nontrivial relative necessity certificate is meaningful.

3.2 SAFE AND EFFICIENT COST CERTIFICATION

Computing an exact deletion cost independently for every feature and candidate group would require many constrained model fits. We therefore solve the shared ambiguity-aware base problems once, derive certified intervals for each deletion cost, and invoke constrained re-optimization only when those intervals do not determine the decision.

Base solutions. We solve the full and feature-deleted instances of equation 4 with an accelerated proximal-gradient method (Beck & Teboulle, 2009), fixing $W_{G,:} = 0$ throughout deleted fits. The loss gradient is evaluated through Euclidean projection onto $\mathcal{Q}_i$, while the $\ell_{2,1}$ term is handled by row-wise group soft thresholding. The intercept-only reference is solved through its reduced gradient problem. Projection and optimization details are given in Appendices D.1 and D.2.

Let $(\widehat{W}, \widehat{b})$ be the full-model solution with objective $\widehat{F}$, and define its stationarity residual by

$$r = \operatorname{dist}\left(0, \partial F(\widehat{W}, \widehat{b})\right). \tag{16}$$

It can be evaluated row-wise from the smooth gradient and the subdifferential of the $\ell_{2,1}$ norm; the explicit formula is given in Appendix D.4. The γ-strong convexity of F gives

$$\widehat{F} - F^\star \le g := \frac{r^2}{2\gamma}, \qquad \|(\widehat{W}, \widehat{b}) - (W^\star, b^\star)\|_2 \le R := \sqrt{\frac{2g}{\gamma}}. \tag{17}$$

Applying the same residual bound to the intercept-only solution gives its returned objective $\widehat{F}_0$ and certified gap g_0. Hence

$$F^\star \in [\widehat{F} - g, \widehat{F}], \qquad F_0^\star \in [\widehat{F}_0 - g_0, \widehat{F}_0]. \tag{18}$$

For the relative tolerance in equation 13, we consequently use

$$\begin{aligned} \epsilon_{\mathrm{L}} &= \eta \max\{0, \widehat{F}_0 - g_0 - \widehat{F}\}, \\ \epsilon_{\mathrm{U}} &= \eta \max\{0, \widehat{F}_0 - \widehat{F} + g\}, \end{aligned} \tag{19}$$

which guarantees $\epsilon \in [\epsilon_{\mathrm{L}}, \epsilon_{\mathrm{U}}]$.

Safe deletion intervals. For any G, the distance bound yields the deletion-cost lower bound

$$\underline{D}(G) := \frac{\gamma}{2}\left(\|\widehat{W}_{G,:}\|_F - R\right)_+^2 \le D(G). \tag{20}$$

An upper bound follows from any feasible predictor that avoids G. In particular, let $\widetilde{W}^G$ be obtained by zeroing the rows of $\widehat{W}$ indexed by G while leaving all other parameters unchanged. Then

$$D(G) \le \overline{D}(G) := F(\widetilde{W}^G, \widehat{b}) - \widehat{F} + g. \tag{21}$$

The feasible objectives for many singleton deletions share the same baseline predictions and can be evaluated in blocks, without running a separate optimizer for every feature.

Combining the tolerance interval in equation 19 with the deletion-cost bounds in equation 20 and equation 21 gives the safe three-way decision

$$\begin{aligned} &\text{non-removable}, && \underline{D}(G) > \epsilon_{\mathrm{U}}, \\ &\text{removable}, && \overline{D}(G) \le \epsilon_{\mathrm{L}}, \\ &\text{unresolved}, && \text{otherwise}. \end{aligned} \tag{22}$$

Only unresolved cases are re-optimized under $W_{G,:} = 0$. If this constrained solver returns objective $\widehat{F}_G$ with certified gap g_G, then

$$\begin{aligned} \underline{D}_{\mathrm{final}}(G) &= \max\left\{\underline{D}(G), 0, \widehat{F}_G - g_G - \widehat{F}\right\}, \\ \overline{D}_{\mathrm{final}}(G) &= \min\left\{\overline{D}(G), \widehat{F}_G - \widehat{F} + g\right\}. \end{aligned} \tag{23}$$

Theorem 2 (Safe deletion-cost certification)**.** *Under $\gamma > 0$ and the certified optimization gaps above, the initial and final deletion intervals contain the exact $D(G)$, and equation 19 contains the exact ϵ. Consequently, every removable or non-removable decision in equation 22 is correct.*

Thus, as in safe screening methods (Fercoq et al., 2015; Ndiaye et al., 2017), screening changes only how many constrained problems are solved; it does not turn an uncertified numerical estimate into a necessity claim. The proof is given in Appendix D.4. In practice, the initial bounds mainly screen clearly removable sets, while borderline non-removability is resolved by constrained refinement.

3.3 Individual and Collective Discovery

Individual necessity. Let $\mathrm{supp}(W) = \{k : \|W_{k,:}\|_2 > 0\}$ denote the active feature support. For a singleton $G = \{k\}$, the equivalence in equation 10 yields the set of individually necessary features

$$\mathcal{N}_\epsilon = \{k : D(\{k\}) > \epsilon\} = \bigcap_{(W,b)\in\mathcal{A}_\epsilon} \mathrm{supp}(W). \tag{24}$$

Hence, an individually necessary feature belongs to the support of every near-optimal predictor, rather than merely the support of the fitted optimum.

Applying equation 22 to every singleton partitions the features into certified necessary, certified removable, and unresolved sets,

$$\begin{aligned} \widehat{\mathcal{N}} &= \{k : \underline{D}(\{k\}) > \epsilon_{\mathrm{U}}\}, \\ \widehat{\mathcal{R}} &= \{k : \overline{D}(\{k\}) \le \epsilon_{\mathrm{L}}\}, \\ \widehat{\mathcal{U}} &= \{1,\dots,d\} \setminus (\widehat{\mathcal{N}} \cup \widehat{\mathcal{R}}). \end{aligned} \tag{25}$$

The bounds guarantee $\widehat{\mathcal{N}} \subseteq \mathcal{N}_\epsilon$ and $\widehat{\mathcal{R}} \cap \mathcal{N}_\epsilon = \varnothing$; unresolved features are not assigned either interpretation without further optimization. Constrained refinement is applied to $\widehat{\mathcal{U}}$ whenever a complete partition is required. For standard feature-budget evaluation, certified necessary features are placed first; features within the necessary and remaining blocks are each ordered by decreasing $\|\widehat{W}_{k,:}\|_2$, with stable index-based tie breaking. This secondary ordering constructs predictive subsets; it is not presented as a necessity certificate.

Collective non-removability. For a larger G, the condition $D(G) > \epsilon$ has a weaker but complementary meaning:

$$\mathrm{supp}(W) \cap G \neq \varnothing, \qquad \forall (W,b) \in \mathcal{A}_\epsilon. \tag{26}$$

We call such a set collectively non-removable. The condition requires every near-optimal model to retain at least one member of G, but does not require any particular member to appear in all models. Collective relevance is especially informative when

$$D(\{k\}) \le \epsilon \quad \text{for every } k \in G, \qquad \text{but} \qquad D(G) > \epsilon. \tag{27}$$

In this regime, every member is individually replaceable, yet the entire group cannot be removed without leaving the near-optimal family.

This definition applies to any feature set G and is independent of how candidate sets are generated. Exhaustively testing all 2^d subsets is nevertheless impossible, so the practical construction below searches a structured family of candidate sets.

The deletion functional is monotone. For any $G_1 \subseteq G_2$,

$$D(G_1) \le D(G_2), \tag{28}$$

because the constraint associated with G_2 removes at least as many feasible predictors as that associated with G_1. This property later enables an efficient search over nested ranked feature sets; its proof and related structural properties appear in Appendix D.3.

The singleton partition determines whether collective discovery is needed. If $\widehat{\mathcal{N}} \neq \varnothing$, the procedure reports the certified individual core and does not invoke a separate group search. If $\widehat{\mathcal{N}} = \varnothing$, however, the absence of an individually necessary feature does not imply that the feature space is

collectively dispensable. We therefore use the collective test as a complementary diagnostic for this case. Collective discovery is performed only after constrained refinement yields a complete singleton partition; if $\widehat{\mathcal{U}} \neq \varnothing$, the unresolved features are reported and no group claim is made. This diagnostic does not enumerate collective structures coexisting with an individual core.

To ensure that a positive group result cannot be inherited from either an individually necessary or an unresolved member, the candidate universe is restricted to $\widehat{\mathcal{R}}$. Let $r_1, \ldots, r_M$ index these certified removable features in non-increasing order of $\|\widehat{W}_{k,:}\|_2$, and construct

$$G_j = \{r_1, \ldots, r_j\}, \qquad j = 1, \ldots, M. \tag{29}$$

These sets are deletion coalitions induced by the predictive ranking, rather than clusters defined by feature correlation or semantics. The first certified G_j satisfying $D(G_j) > \epsilon$ identifies the smallest prefix for which individually replaceable features become collectively non-removable. Here, "smallest" is prefix-relative, not global.

Because $D(G_j)$ is monotone in j, exponential search first brackets a transition and binary refinement narrows it. With exact decisions this finds the first non-removable prefix using logarithmically many group tests after bracketing. With interval decisions, an unresolved prefix is never treated as removable; the procedure reports a certified removable/non-removable bracket and its conservative non-removable endpoint. Candidate sets are examined only up to a prescribed feature budget, so failure to find a crossing is reported as no certified collective set within the search range, rather than as evidence that none exists. Appendix D.5 establishes these claims. Algorithm 1 in Appendix D.6 summarizes the complete certification and structured-search procedure.

4 Experiments

We evaluate the framework structurally and predictively. Controlled experiments first test recovery against known feature roles; real PML benchmarks then assess the prevalence and predictive consequences of the identified structures, followed by sensitivity analysis to the near-optimal tolerance.

4.1 Experimental Protocol

Our primary comparison covers five controlled benchmark families (Birds, Corel5k, Enron, HumanPseAAC, and Mediamill) and three naturally ambiguous PML datasets (Music emotion, Music style, and Mirflickr). We compare against eight representative PML and PML feature-selection methods under a common one-vs-rest linear-SVM evaluation. Following the shared protocol, we evaluate the top 1%–20% of each feature ranking and report Ranking Loss, Coverage Error, Average Precision, and Micro-F1. Hyperparameters are selected on a fixed validation partition, and target labels are never used by the feature-selection objective. Unless otherwise stated, $\eta = 0.01$. Dataset variants, parameter grids, preprocessing, and the complete protocol are provided in Appendix B. The formulation covers CML through $s_i = \mathbf{1} - c_i$, whereas our experiments focus on PML.

4.2 Controlled Structural Validation

To verify whether our method identifies the intended relevance structures, we first conduct experiments on controlled synthetic data with known feature roles. Each dataset contains $2{,}000$ examples, 20 features, and eight labels, exactly two of which are relevant to each example. The observed set s_i contains these two labels together with zero, two, or four randomly sampled irrelevant labels; hence $|s_i| \in \{2, 4, 6\}$ controls the degree of label ambiguity. We generate 30 datasets with independent random seeds for every candidate-set size and scenario.

The two scenarios encode different ground-truth deletion outcomes. In the *individual* scenario, a unique feature x_u carries the predictive signal and the other 19 features are independent Gaussian noise. Deleting x_u should therefore preclude a near-optimal explanation, making it necessary, whereas deleting any noise feature should remain feasible. In the *collective* scenario, the signal is duplicated into two features x_a and x_b. Either singleton can be deleted because the other copy remains, but deleting $\{x_a, x_b\}$ removes the signal altogether. The expected outcome is thus that both singletons are removable while their pair is non-removable; the remaining 18 noise features should

again be removable. These prescribed outcomes are never supplied to the certification or search procedure and are used only for evaluation.

Table 1: Recovery of known feature structures. An entry a/b means that a of the b tested deletion objects received their prescribed outcome. For example, $570 = 19 \times 30$ noise singletons in the individual scenario, whereas $60 = 2 \times 30$ signal-copy singletons in the collective scenario.

Scenario	Deleted object	Expected outcome	$\lvert s_i\rvert = 2$	$\lvert s_i\rvert = 4$	$\lvert s_i\rvert = 6$
Individual	Unique feature	Necessary	30/30	30/30	30/30
Individual	Noise singleton	Removable	570/570	570/570	570/570
Collective	Either signal copy	Removable	60/60	60/60	60/60
Collective	Both signal copies	Non-removable	30/30	30/30	30/30
Collective	Noise singleton	Removable	540/540	540/540	540/540
Collective	First discovered set	Exactly the signal pair	30/30	30/30	30/30

Table 1 shows that every prescribed singleton outcome is recovered without an unresolved decision. In the collective scenario, the search returns $\{x_a, x_b\}$ as the first non-removable set in all 90 runs, without including noise. The same outcomes hold for $|s_i| = 2, 4, 6$, showing that the distinction between individual necessity and collective non-removability persists as label ambiguity increases.

4.3 Evidence on Real Data

We next examine whether the two relevance structures identified above also arise on real benchmark problems. Among the 18 settings reported below, 14 contain a nonempty certified individual core. The remaining four admit a complete singleton-removability certificate and therefore enter the collective search; each yields a certified non-removable prefix. Thus the reported benchmarks contain both individually indispensable features and sets whose relevance is visible only collectively.

Table 2 evaluates the predictive utility of the resulting ranking using Ranking Loss, where lower values are better. Each entry reports the mean and budget-wise standard deviation over feature budgets from 1% to 20%; controlled settings are grouped by candidate-set size c, followed by the real PML datasets. Complementary metrics are reported in Appendix C.

Table 2: Ranking Loss (mean $\pm$ budget-wise standard deviation) over feature budgets from 1% to 20%. Lower is better. Best results are shown in bold.

Dataset	FIND	PML-FSLA	PML-FSMIR	PML-FSSO	fPML	PML-LD	PML-FP	PML-MAP	PAMB
Birds ($c = 3$)	**.114±.005**	.275±.056	.355±.101	.292±.056	.354±.070	.327±.006	.910±.047	.564±.091	.557±.107
Birds ($c = 4$)	**.114±.006**	.275±.057	.406±.138	.304±.055	.349±.054	.445±.124	.923±.040	.554±.087	.501±.098
Birds ($c = 5$)	**.128±.004**	.273±.057	.361±.094	.307±.058	.384±.060	.350±.140	.916±.035	.560±.086	.564±.101
Corel5k ($c = 3$)	**.177±.004**	.425±.159	.532±.190	.590±.189	.737±.148	.816±.122	.846±.094	.660±.141	.848±.100
Corel5k ($c = 4$)	**.176±.006**	.432±.159	.548±.182	.614±.176	.761±.137	.830±.118	.846±.096	.663±.147	.869±.103
Corel5k ($c = 5$)	**.183±.008**	.430±.160	.571±.182	.627±.176	.779±.125	.855±.120	.858±.085	.679±.137	.803±.136
Enron ($c = 5$)	**.245±.007**	.354±.106	.493±.095	.377±.078	.369±.058	.417±.091	.950±.005	.638±.099	.496±.146
Enron ($c = 7$)	**.256±.011**	.355±.073	.493±.131	.392±.055	.436±.053	.416±.201	.954±.006	.560±.103	.455±.170
Enron ($c = 9$)	**.279±.011**	.357±.100	.548±.118	.399±.047	.463±.048	.484±.130	.953±.009	.572±.084	.552±.159
HumanPseAAC ($c = 3$)	**.196±.010**	.244±.000	.285±.045	.261±.008	.273±.022	.217±.059	.927±.049	.439±.013	.395±.069
HumanPseAAC ($c = 4$)	**.200±.011**	.244±.000	.326±.056	.260±.012	.287±.024	.265±.020	.926±.049	.436±.025	.702±.080
HumanPseAAC ($c = 5$)	.224±.013	.244±.000	.316±.019	.284±.012	.330±.020	**.202±.043**	.927±.048	.506±.011	.594±.057
Mediamill ($c = 3$)	**.103±.006**	.142±.001	.202±.089	.146±.004	.561±.032	.170±.031	.495±.071	.608±.046	.182±.056
Mediamill ($c = 4$)	**.100±.005**	.142±.000	.162±.032	.147±.007	.547±.037	.168±.033	.338±.141	.612±.020	.201±.097
Mediamill ($c = 5$)	**.106±.006**	.143±.001	.171±.023	.147±.004	.546±.026	.243±.053	.316±.130	.527±.020	.233±.126
Music emotion	**.271±.018**	.360±.000	.489±.152	.370±.015	.347±.050	.364±.017	.998±.001	.452±.041	.371±.070
Music style	**.159±.013**	.227±.000	.505±.207	.241±.009	.368±.069	.314±.059	.997±.001	.507±.097	.512±.070
Mirflickr	**.130±.021**	.252±.000	.297±.059	.265±.022	.256±.037	.300±.080	.410±.087	.346±.082	.149±.055

Across the 18 settings, our method achieves the best Ranking Loss in 17; the exception is HumanPseAAC ($c = 5$), where it ranks second to PML-LD. The advantage spans controlled and real PML benchmarks, indicating that the learned ordering is not tied to one ambiguity construction. The complementary metrics show the same pattern: our method ranks first on all 18 settings for Coverage Error and Micro-F1 and on 16 for Average Precision. Appendix C provides a more detailed result include cross-metric and budget-wise analysis.

Matched-deletion validation. Finally, we test whether the certificates identify behaviorally consequential objects rather than descriptive structures. At matched budgets, we compare deleting a certified object with deleting an equally sized random control, separately for individual cores and collective sets. For the settings in Table 2, we average ten controls within each fold and then aggregate equally across budgets, ambiguity settings, and dataset families; thus no family receives extra weight from having more observations. Valid targets yield individual-core comparisons for 11 settings from six families and collective-set comparisons for five settings from three families. Table 3 standardizes metric directions so that positive degradation is worse; its positive certified-minus-random gaps show that certified objects are harder to replace than size-matched controls.

Table 3: Macro-averaged performance degradation ($\times 10^{-3}$) after deleting a certified object or an equally sized random control. Gap is Certified minus Random; a positive gap favors the certificate.

	Individual core ($N = 6$ families)			Collective set ($N = 3$ families)		
Metric	Certified	Random	Gap	Certified	Random	Gap
Ranking Loss	20.49	2.99	17.50	3.14	−0.22	3.36
Coverage Error	22.36	3.81	18.55	3.36	−0.29	3.65
Average Precision	18.57	5.22	13.35	4.47	0.16	4.31
Micro-F1	21.36	3.49	17.87	8.15	2.10	6.05

4.4 Sensitivity to the Near-Optimal Tolerance

Unlike the training parameters λ and γ, η changes which predictors count as near-optimal without changing the fitted objective. We therefore re-threshold the saved singleton deletion-cost intervals at $\eta \in \{0.001, 0.0025, 0.005, 0.01, 0.02, 0.05, 0.1\}$, without refitting. Figure 2 averages the resulting feature proportions over 18 settings.

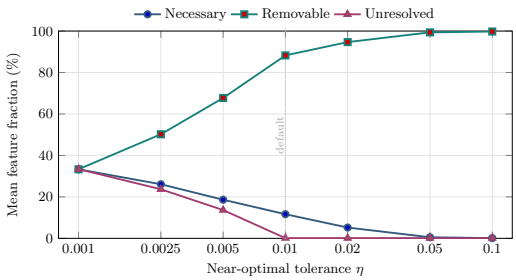


Figure 2: Sensitivity of the certified singleton partition to the near-optimal tolerance. The vertical line marks the default $\eta = 0.01$.

As predicted, expanding the admissible family decreases the necessary fraction from 33.3% to 0.1% and raises the removable fraction from 33.3% to 99.8%. At the default $\eta = 0.01$, the proportions are 11.7% necessary, 88.2% removable, and 0.1% unresolved; 14 settings contain an individual core and four yield a collective set. Thus, the default preserves both structures and an almost complete safe partition. The monotone shift is expected: a larger tolerance admits more alternatives after deletion. Because the saved intervals are not refined separately for each η, unresolved rates at small tolerances may be conservative.

5 Conclusion

We distinguish selecting one sparse support from certifying which features remain unavoidable across near-optimal predictors under inexact multi-label supervision. FIND formalizes this distinction through counterfactual deletion: the resulting deletion cost characterizes whether a feature or feature set can be avoided within a prescribed tolerance, yielding individual necessity and collective non-removability, while safe bounds enable efficient certification.Experiments validate both structures while showing that the resulting rankings retain strong predictive performance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and proofreading, including checking grammar, spelling, notation, and typographical errors. We also used generative AI tools during the preliminary stage of the study to support literature exploration and retrieval. All references and literature identified with AI assistance were independently verified against their original sources, and all AI-assisted revisions were reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code used in this work is available in our anonymous repository at `https://anonymous.4open.science/r/find-5172/`. The appendix provides complete experimental details, including dataset descriptions, preprocessing, hyperparameter settings, and evaluation protocols, as well as additional results. It also includes optimization details, theoretical assumptions, complete proofs, and the full certification and discovery procedure. Together with the released code, these materials support reproduction of the results reported in this work.

REFERENCES

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Ali Braytee, Wei Liu, Daniel R Catchpoole, and Paul J Kennedy. Multi-label feature selection using correlation information. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1649–1656, 2017.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Marc-Etienne Brunet, Ashton Anderson, and Richard Zemel. Implications of model indeterminacy for explanations of automated decisions. *Advances in Neural Information Processing Systems*, 35:7810–7823, 2022.

Rudin Cynthia, Zhong Chudi, Semenova Lesia, Seltzer Margo, Parr Ronald, Liu Jiachang, Katta Srikar, Donnelly Jon, Chen Harry, and Boner Zachery. Position: amazing things come from having many good models. In *Proceedings of the 41st International Conference on Machine Learning, ICML*, volume 24, 2024.

Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.

Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward Browne. The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. *Advances in Neural Information Processing Systems*, 36:6267–6279, 2023.

Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European conference on computer vision*, pp. 97–112. Springer, 2002.

André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. *Advances in neural information processing systems*, 14, 2001.

Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the lasso. In *International conference on machine learning*, pp. 333–342. PMLR, 2015.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of machine learning research*, 20(177):1–81, 2019.

Wanfu Gao, Liang Hu, and Ping Zhang. Class-specific mutual information variation for feature selection. *Pattern Recognition*, 79:328–339, 2018.

Wanfu Gao, Yonghao Li, and Liang Hu. Multilabel feature selection with constrained latent structure shared term. *IEEE transactions on neural networks and learning systems*, 34(3):1253–1262, 2021.

Wanfu Gao, Pingting Hao, Yang Wu, and Ping Zhang. A unified low-order information-theoretic feature selection framework for multi-label learning. *Pattern Recognition*, 134:109111, 2023.

Wanfu Gao, Hanlin Pan, Qingqi Han, and Kunpeng Liu. Noise-resistant label reconstruction feature selection for partial multi-label learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/576.

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th annual conference on learning theory*, pp. 341–358. JMLR Workshop and Conference Proceedings, 2011.

Xiuwen Gong, Dong Yuan, and Wei Bao. Understanding partial multi-label learning via mutual information. *Advances in Neural Information Processing Systems*, 34:4147–4156, 2021.

Jorge Gonzalez-Lopez, Sebastián Ventura, and Alberto Cano. Distributed multi-label feature selection using individual mutual information measures. *Knowledge-Based Systems*, 188:105052, 2020.

Pingting Hao, Liang Hu, and Wanfu Gao. Partial multi-label feature selection via subspace optimization. *Information Sciences*, 648:119556, 2023.

Juncheng Hu, Yonghao Li, Wanfu Gao, and Ping Zhang. Robust multi-label feature selection with dual-graph regularization. *Knowledge-Based Systems*, 203:106126, 2020a.

Liang Hu, Yonghao Li, Wanfu Gao, Ping Zhang, and Juncheng Hu. Multi-label feature selection with shared common mode. *Pattern Recognition*, 104:107344, 2020b.

Ling Jian, Jundong Li, Kai Shu, and Huan Liu. Multi-label informed feature selection. In *Ijcai*, volume 16, pp. 1627–33, 2016.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research*, 24(364):1–50, 2023.

Yonghao Li, Liang Hu, and Wanfu Gao. Robust sparse and low-redundancy multi-label feature selection with dynamic local and global structure preservation. *Pattern Recognition*, 134:109120, 2023a.

Yonghao Li, Liang Hu, and Wanfu Gao. Multi-label feature selection via robust flexible sparse regularization. *Pattern Recognition*, 134:109074, 2023b.

Ziwei Li, Gengyu Lyu, and Songhe Feng. Partial multi-label learning via multi-subspace representation. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pp. 2612–2618, 2021.

Yaojin Lin, Qinghua Hu, Jinghua Liu, and Jie Duan. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, 168:92–103, 2015.

Bing-Qing Liu, Bin-Bin Jia, and Min-Ling Zhang. Towards enabling binary decomposition for partial multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 45 (11):13203–13217, 2023.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, pp. 6765–6774. PMLR, 2020.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017.

Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. *Advances in neural information processing systems*, 23, 2010.

Hanlin Pan, Kunpeng Liu, and Wanfu Gao. Reconsidering feature structure information and latent space alignment in partial multi-label feature selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19786–19794, 2025.

Thuan Pham, Xiaohui Tao, Ji Zhang, Jianming Yong, Yuefeng Li, and Haoran Xie. Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-based systems*, 235:107662, 2022.

Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5016–5023, 2019.

Lijuan Sun, Songhe Feng, Jun Liu, Gengyu Lyu, and Congyan Lang. Global-local label correlation for partial multi-label learning. *IEEE Transactions on Multimedia*, 24:581–593, 2021.

Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

Haobo Wang, Weiwei Liu, Yang Zhao, Chen Zhang, Tianlei Hu, and Gang Chen. Discriminative and correlative partial multi-label learning. In *IJCAI*, pp. 3691–3697, 2019.

Wei Wang, Tianhao Ma, Ming-Kun Xie, Gang Niu, and Masashi Sugiyama. Rethinking consistent multi-label classification under inexact supervision. In *International Conference on Learning Representations*, volume 2026, pp. 102638–102663, 2026.

Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *international conference on machine learning*, pp. 3780–3788. PMLR, 2017.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Ming-Kun Xie and Sheng-Jun Huang. Semi-supervised partial multi-label learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 691–700. IEEE, 2020.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2021.

Ning Xu, Yun-Peng Liu, and Xin Geng. Partial multi-label learning with label distribution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6510–6517, 2020.

Yan Yan and Yuhong Guo. Adversarial partial multi-label learning with label disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10568–10576, 2021.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *2018 IEEE international conference on data mining (ICDM)*, pp. 1398–1403. IEEE, 2018.

Tingting Yu, Guoxian Yu, Jun Wang, and Maozu Guo. Partial multi-label learning with label and feature collaboration. In *International conference on database systems for advanced applications*, pp. 621–637. Springer, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Min-Ling Zhang and Jun-Peng Fang. Partial multi-label learning via credible label elicitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3587–3599, 2020.

Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048, 2007.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

Ping Zhang and Wanfu Gao. Feature relevance term variation for multi-label feature selection. *Applied Intelligence*, 51(7):5095–5110, 2021.

Ping Zhang, Guixia Liu, Wanfu Gao, and Jiazhi Song. Multi-label feature selection considering label supplementation. *Pattern recognition*, 120:108137, 2021.

A Related Work

Learning with inexact multi-label supervision. Multi-label learning extends classification to outputs containing several interdependent labels, motivating methods and analyses for neighborhood prediction, kernel ranking, consistency, and evaluation measures (Zhang & Zhou, 2007; Elisseeff & Weston, 2001; Gao & Zhou, 2011; Wu & Zhou, 2017; Zhang & Zhou, 2013). PML weakens the exact-label assumption by providing a candidate set. Representative approaches disambiguate candidates through feature-induced representations, low-rank or sparse structure, label correlations, label distributions, label–feature collaboration, noisy-label identification, multiple subspaces, adversarial learning, or information-theoretic criteria (Xie & Huang, 2018; Yu et al., 2018; Sun et al., 2019; Wang et al., 2019; Zhang & Fang, 2020; Xu et al., 2020; Yu et al., 2020; Xie & Huang, 2021; Li et al., 2021; Yan & Guo, 2021; Gong et al., 2021). Related extensions exploit unlabeled instances or global–local label correlations, while recent work unifies partial and complementary annotations as inexact supervision (Xie & Huang, 2020; Sun et al., 2021; Wang et al., 2026).

Feature selection for multi-label learning. Multi-label feature selection commonly learns a sparse shared representation or evaluates features using label-dependent relevance and redundancy (Nie et al., 2010; Jian et al., 2016; Lin et al., 2015; Braytee et al., 2017; Gonzalez-Lopez et al., 2020). Subsequent methods incorporate class-specific information, dual graphs, shared modes, constrained latent structures, low-order information, flexible sparsity, dynamic local–global structure, or label supplementation (Gao et al., 2018; Hu et al., 2020a;b; Gao et al., 2021; 2023; Li et al., 2023b;a; Zhang & Gao, 2021; Zhang et al., 2021). PML-specific selectors further couple feature sparsity with credible-label reconstruction and latent-space structure (Hao et al., 2023; Pan et al., 2025; Gao et al., 2025). These methods are designed to recover a useful support from a fitted solution; our concern is instead whether deleting a feature set leaves any comparably good admissible solution.

Near-optimal model families and stable relevance. The gap between a fitted model and other comparably accurate explanations echoes the longstanding distinction between prediction and model-based interpretation (Breiman, 2001). Model class reliance evaluates variable importance over a class of competitive models (Fisher et al., 2019), while predictive multiplicity measures prediction disagreement among similarly accurate classifiers (Marx et al., 2020). Later studies trace variable importance across the cloud of good models (Dong & Rudin, 2020), summarize its variation as a distribution (Donnelly et al., 2023), or seek consensus among feature-attribution rankings (Laberge et al., 2023); model indeterminacy can likewise alter downstream inference (Brunet et al., 2022). Collectively, these works motivate analyzing many good models (Cynthia et al., 2024), but generally treat supervision as fixed. Our setting couples alternative target assignments with alternative predictive supports. We therefore re-optimize after deleting a feature set and test whether any admissible near-optimal predictor can avoid it.

B Experimental Details

The primary predictive comparison contains five controlled multi-label dataset families, Birds, Corel5k, Enron, HumanPseAAC, and Mediamill, and three naturally ambiguous PML datasets, Music emotion, Music style, and Mirflickr. For each controlled family, exact labels are retained only for evaluation and augmented with irrelevant labels to form candidate sets. We use $c \in \{3, 4, 5\}$ for all controlled families except Enron, for which $c \in \{5, 7, 9\}$, giving 18 primary settings.

Tables 4 and 5 summarize the characteristics of the controlled and naturally ambiguous PML datasets, respectively. Here, avg.#TLs and avg.#CLs denote the average numbers of true and candi-

Table 4: Characteristics of the controlled multi-label datasets used in the experiments.

Datasets	Domain	#Ex.	#Feat.	#Labels	avg.#TLs	c settings
Birds	Audio	645	152	20	1.49	3, 4, 5
Corel5k	Image	3,343	499	15	1.58	3, 4, 5
Enron	Text	856	53	15	2.42	5, 7, 9
HumanPseAAC	Biology	3,106	440	14	1.19	3, 4, 5
Mediamill	Video	8,276	120	15	1.05	3, 4, 5

date labels per instance, respectively, while c denotes the candidate-set size used for the controlled ambiguity settings.

We compare with PML-FSSO (Hao et al., 2023), PML-FSLA (Pan et al., 2025), PML-FSMIR (Gao et al., 2025), fPML (Yu et al., 2018), PML-LD (Xu et al., 2020), PML-FP (Xie & Huang, 2018), PML-MAP (Zhang & Fang, 2020), and PAMB (Liu et al., 2023). All feature rankings are evaluated with the same one-vs-rest linear SVM. Dedicated feature-selection baselines use the rankings returned by their implementations. For general PML learners, we stack their label-wise linear coefficient vectors into a shared coefficient matrix and rank features by its row-wise ℓ_2 norm; thus no target labels or external filter are used to construct their rankings. We search $\lambda/\lambda_{\max} \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ and $\gamma/L \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, expanding the grid when the best configuration lies on its boundary.

We construct five fixed partitions: one is reserved for hyperparameter selection using clean validation targets and the remaining four provide held-out evaluation. Preprocessing is fitted on the training portion of each evaluation fold. The dimensions in Tables 4 and 5 refer to the supplied preprocessed benchmark matrices; FIND performs no additional instance, feature, or label pruning. Within each fold, feature standardization is fitted using only the training partition. Certified necessary features precede the remaining features; both blocks are ordered by decreasing $\|\widehat{W}_{k,:}\|_2$ with stable index-based tie breaking. The selected features are passed to a one-vs-rest linear SVM trained on the observed partial-label matrix and evaluated against held-out targets. Budget positions are averaged within each fold before fold-level aggregation, and ambiguity settings from the same dataset family are not treated as independent samples in statistical tests. The appendix reports the other three complementary metrics stated in Section 4.1; individual budget positions are reserved for the complete result archive.

Synthetic construction. For each run, draw a balanced regime variable $h_i \in \{+1, -1\}$ and define $z_i = 2h_i + 0.2\xi_i$ with $\xi_i \sim \mathcal{N}(0, 1)$. The exact target contains labels $\{1, 2\}$ when $h_i = +1$ and labels $\{3, 4\}$ when $h_i = -1$; the remaining labels are irrelevant. All nonsignal features are sampled independently from $\mathcal{N}(0, 1)$. In the individual scenario, $x_{iu} = z_i$ for one feature u. In the collective scenario, two features are exact copies, $x_{ia} = x_{ib} = z_i$, so either copy can substitute for the other but deleting both removes the signal. The observed candidate set is formed by adding $d_c \in \{0, 2, 4\}$ labels sampled uniformly without replacement from the six irrelevant labels, giving $|s_i| = 2 + d_c$. Features are standardized after generation, and neither the regime nor the prescribed feature roles are supplied to the learner.

Matched-deletion protocol. Certificates and removal targets are determined separately within each evaluation fold. At budgets 5%, 10%, and 20%, we remove the certified target from the top-k subset, fill the vacated positions with the next ranked features, and retrain the same one-vs-rest linear SVM; each random control removes an equally sized, disjoint subset of the original top-k features and is processed identically. A fold is omitted when its target is absent, exceeds k, or leaves too few eligible control features. The individual comparison uses Birds ($c = 3, 4, 5$), Corel5k ($c = 3, 4, 5$), HumanPseAAC ($c = 3, 4$), Music emotion, Music style, and Mirflickr; the collective comparison uses Corel5k ($c = 3$), HumanPseAAC ($c = 5$), and Mediamill ($c = 3, 4, 5$). Corel5k ($c = 3$) contributes to both because its fold-specific certificates exhibit different valid target types. This fold-level construction explains why these counts need not match the one-certificate-per-setting summary in the main text.

Table 5: Characteristics of the real-world PML datasets used in the experiments.

Datasets	Domain	#Ex.	#Feat.	#Labels	avg.#CLs	avg.#TLs
Mirflickr	Image	10,433	100	7	3.35	1.77
Music emotion	Audio	6,833	98	11	5.29	2.42
Music style	Audio	6,839	98	10	6.04	1.44

Table 6: Coverage Error ↓ (mean ± budget-wise standard deviation) over feature budgets from 1% to 20%. Best results are shown in bold.

Dataset	FIND	PML-FSLA	PML-FSMIR	PML-FSSO	fPML	PML-LD	PML-FP	PML-MAP	PAMB
Birds ($c = 3$)	**.370±.020**	.434±.011	.451±.016	.421±.015	.461±.024	.589±.007	.546±.020	.644±.020	.732±.065
Birds ($c = 4$)	**.386±.019**	.429±.013	.487±.005	.440±.012	.453±.009	.628±.052	.527±.008	.636±.013	.662±.081
Birds ($c = 5$)	**.420±.011**	.433±.013	.458±.012	.458±.019	.485±.014	.505±.081	.527±.009	.626±.012	.730±.075
Corel5k ($c = 3$)	**.237±.009**	.330±.041	.362±.051	.424±.038	.468±.008	.522±.056	.471±.002	.479±.007	.543±.009
Corel5k ($c = 4$)	**.236±.010**	.332±.040	.371±.044	.434±.029	.469±.008	.526±.052	.475±.003	.492±.006	.545±.015
Corel5k ($c = 5$)	**.244±.013**	.330±.041	.399±.033	.445±.022	.470±.004	.489±.042	.476±.004	.495±.007	.539±.034
Enron ($c = 5$)	**.322±.011**	.407±.057	.547±.046	.443±.045	.437±.024	.514±.029	.717±.001	.634±.030	.561±.032
Enron ($c = 7$)	**.340±.016**	.406±.037	.546±.053	.477±.033	.523±.018	.475±.031	.717±.001	.588±.033	.470±.023
Enron ($c = 9$)	**.374±.020**	.407±.052	.576±.037	.490±.021	.555±.014	.508±.037	.717±.001	.619±.028	.527±.044
HumanPseAAC ($c = 3$)	**.214±.011**	.249±.000	.289±.033	.268±.007	.280±.020	.259±.053	.512±.006	.436±.010	.367±.050
HumanPseAAC ($c = 4$)	**.218±.013**	.249±.000	.323±.030	.266±.011	.293±.023	.294±.018	.510±.007	.438±.018	.561±.067
HumanPseAAC ($c = 5$)	**.243±.016**	.249±.000	.323±.018	.289±.011	.336±.018	.244±.041	.511±.007	.503±.009	.474±.044
Mediamill ($c = 3$)	**.108±.007**	.142±.001	.197±.083	.144±.003	.531±.030	.162±.029	.471±.067	.575±.043	.172±.051
Mediamill ($c = 4$)	**.106±.005**	.142±.000	.161±.030	.145±.006	.519±.035	.157±.031	.325±.131	.578±.019	.196±.089
Mediamill ($c = 5$)	**.112±.006**	.143±.001	.168±.022	.146±.004	.517±.024	.226±.048	.305±.121	.500±.019	.226±.117
Music emotion	**.382±.021**	.566±.001	.505±.031	.501±.006	.500±.018	.580±.013	.589±.000	.582±.005	.534±.063
Music style	**.209±.017**	.296±.000	.448±.086	.309±.007	.404±.062	.360±.045	.638±.000	.503±.036	.539±.065
Mirflickr	**.157±.024**	.382±.000	.357±.013	.339±.014	.313±.019	.273±.069	.458±.043	.384±.036	.355±.078

C Additional Predictive Results

This section reports the complete comparison under Coverage Error, Average Precision, and Micro-F1. As in the main table, every entry is the mean and budget-wise standard deviation over all 20 feature budgets from 1% to 20%; this dispersion describes the budget trajectory rather than fold-level statistical uncertainty.

Taken together with Table 2, the four metrics yield 72 dataset–metric comparisons. The proposed method ranks first in 69 of them: 17/18 for Ranking Loss, 18/18 for Coverage Error, 16/18 for Average Precision, and 18/18 for Micro-F1. It also outperforms the eight baselines in 573 of the 576 direct comparisons (99.5%). This agreement is important because the metrics emphasize different aspects of multi-label prediction: Ranking Loss and Coverage Error evaluate label ordering, whereas Average Precision and Micro-F1 measure retrieval and classification quality. Improvements in all four directions therefore cannot be attributed to optimizing a single evaluation criterion.

Micro-F1 binarizes every one-vs-rest LinearSVC score at the common zero threshold. Under severe label imbalance, this can yield no positive predictions and hence a zero score even when the continuous ranking remains informative. Because every method uses the same classifier and threshold, the comparison is controlled; nevertheless, Ranking Loss and Average Precision provide the more direct evidence of ranking quality.

The pattern is also consistent across the two benchmark types. On the 15 controlled settings, our method is best in 58 of 60 dataset–metric comparisons; on the three real-world PML datasets, it is best in 11 of 12. The three exceptions are narrow and localized. PML-LD has lower Ranking Loss on HumanPseAAC ($c = 5$), .202 versus .224; PML-FSLA has slightly higher Average Precision on Enron ($c = 9$), .426 versus .424; and PAMB leads Average Precision on Mirflickr, .821 versus .805. In each case, our method remains competitive on that metric and leads the other reported criteria for the same setting. Thus, the aggregate advantage does not depend exclusively on either a controlled ambiguity mechanism or the three naturally ambiguous datasets.

Figures 3–6 show the complete budget trajectories for all 18 primary settings: the three real-world PML datasets and all ambiguity settings from the five controlled families. Fixing the panels across metrics makes the trends directly comparable and shows whether an advantage persists beyond isolated budgets. Across the resulting 360 budget positions per metric, our method is best at 90.0%

Table 7: Average Precision ↑ (mean ± budget-wise standard deviation) over feature budgets from 1% to 20%. Best results are shown in bold.

Dataset	FIND	PML-FSLA	PML-FSMIR	PML-FSSO	fPML	PML-LD	PML-FP	PML-MAP	PAMB
Birds ($c = 3$)	**.638±.009**	.513±.035	.532±.054	.573±.034	.522±.049	.602±.016	.198±.017	.265±.038	.261±.099
Birds ($c = 4$)	**.635±.013**	.514±.036	.485±.074	.556±.034	.527±.036	.530±.095	.188±.010	.272±.045	.274±.078
Birds ($c = 5$)	**.608±.006**	.513±.036	.510±.046	.562±.034	.495±.037	.569±.088	.190±.009	.272±.037	.227±.090
Corel5k ($c = 3$)	**.564±.018**	.424±.058	.380±.069	.377±.074	.309±.057	.337±.086	.237±.025	.335±.050	.237±.009
Corel5k ($c = 4$)	**.562±.016**	.419±.059	.375±.066	.361±.065	.289±.050	.325±.080	.232±.022	.312±.045	.234±.008
Corel5k ($c = 5$)	**.554±.011**	.425±.061	.365±.063	.333±.059	.272±.039	.323±.066	.226±.015	.301±.038	.266±.048
Enron ($c = 5$)	**.458±.010**	.438±.014	.393±.003	.421±.010	.414±.005	.409±.015	.391±.001	.342±.010	.433±.017
Enron ($c = 7$)	**.460±.009**	.427±.003	.379±.009	.409±.006	.393±.004	.435±.017	.391±.002	.365±.003	.407±.019
Enron ($c = 9$)	.424±.008	**.426±.006**	.364±.008	.391±.004	.369±.007	.392±.028	.390±.002	.353±.005	.360±.009
HumanPseAAC ($c = 3$)	**.559±.003**	.444±.000	.402±.032	.382±.020	.395±.017	.500±.053	.222±.010	.296±.009	.396±.044
HumanPseAAC ($c = 4$)	**.553±.007**	.444±.000	.394±.028	.390±.017	.395±.022	.486±.019	.220±.011	.283±.018	.321±.041
HumanPseAAC ($c = 5$)	**.530±.007**	.444±.000	.388±.019	.384±.016	.370±.025	.522±.054	.220±.009	.257±.009	.385±.035
Mediamill ($c = 3$)	**.738±.016**	.616±.001	.614±.042	.615±.001	.353±.024	.557±.081	.364±.013	.177±.015	.547±.112
Mediamill ($c = 4$)	**.739±.015**	.616±.000	.609±.045	.614±.002	.363±.021	.548±.077	.445±.060	.146±.005	.567±.143
Mediamill ($c = 5$)	**.735±.013**	.610±.007	.583±.024	.614±.001	.368±.011	.532±.095	.454±.063	.186±.016	.552±.164
Music emotion	**.580±.020**	.496±.000	.530±.045	.515±.002	.536±.019	.529±.025	.343±.000	.398±.008	.517±.082
Music style	**.709±.013**	.652±.000	.457±.054	.646±.006	.426±.072	.458±.100	.380±.001	.320±.012	.322±.049
Mirflickr	.805±.017	.682±.000	.619±.039	.622±.025	.650±.030	.585±.068	.500±.051	.593±.041	**.821±.045**

Table 8: Micro-F1 ↑ (mean ± budget-wise standard deviation) over feature budgets from 1% to 20%. Best results are shown in bold.

Dataset	FIND	PML-FSLA	PML-FSMIR	PML-FSSO	fPML	PML-LD	PML-FP	PML-MAP	PAMB
Birds ($c = 3$)	**.448±.024**	.000±.000	.193±.069	.318±.108	.396±.049	.367±.121	.000±.000	.223±.084	.134±.098
Birds ($c = 4$)	**.457±.034**	.000±.000	.012±.029	.321±.100	.358±.073	.293±.147	.000±.000	.210±.018	.089±.041
Birds ($c = 5$)	**.427±.026**	.000±.000	.024±.038	.375±.047	.362±.060	.332±.146	.000±.000	.131±.023	.070±.053
Corel5k ($c = 3$)	**.242±.075**	.000±.000	.199±.051	.013±.021	.000±.000	.154±.073	.000±.000	.074±.044	.053±.026
Corel5k ($c = 4$)	**.270±.089**	.103±.051	.199±.098	.010±.031	.000±.000	.125±.055	.000±.000	.040±.046	.045±.024
Corel5k ($c = 5$)	**.354±.026**	.128±.064	.120±.033	.074±.020	.014±.018	.083±.038	.000±.000	.059±.046	.051±.025
Enron ($c = 5$)	**.441±.046**	.207±.025	.014±.020	.000±.000	.000±.000	.041±.047	.000±.000	.144±.038	.104±.064
Enron ($c = 7$)	**.442±.020**	.000±.000	.102±.029	.016±.019	.021±.040	.046±.049	.000±.000	.153±.044	.201±.155
Enron ($c = 9$)	**.282±.009**	.000±.000	.052±.010	.039±.027	.065±.004	.011±.019	.000±.000	.108±.022	.030±.024
HumanPseAAC ($c = 3$)	**.272±.043**	.178±.005	.000±.000	.000±.000	.083±.045	.073±.049	.000±.000	.211±.027	.035±.030
HumanPseAAC ($c = 4$)	**.332±.024**	.178±.006	.000±.000	.000±.000	.072±.034	.066±.039	.000±.000	.215±.024	.009±.012
HumanPseAAC ($c = 5$)	**.355±.004**	.177±.005	.061±.033	.044±.023	.052±.033	.023±.016	.000±.000	.132±.014	.030±.030
Mediamill ($c = 3$)	**.604±.028**	.000±.000	.000±.000	.000±.000	.450±.128	.000±.000	.355±.079	.123±.031	.000±.000
Mediamill ($c = 4$)	**.616±.023**	.000±.000	.000±.000	.000±.000	.490±.045	.000±.000	.458±.071	.078±.020	.000±.000
Mediamill ($c = 5$)	**.603±.019**	.000±.000	.000±.000	.000±.000	.466±.036	.000±.000	.402±.058	.112±.041	.014±.024
Music emotion	**.488±.012**	.000±.000	.000±.000	.062±.074	.023±.033	.019±.010	.000±.000	.240±.077	.040±.021
Music style	**.266±.005**	.000±.000	.000±.000	.155±.196	.000±.000	.004±.002	.000±.000	.144±.064	.020±.007
Mirflickr	**.681±.013**	.449±.022	.287±.042	.000±.000	.270±.077	.045±.000	.000±.000	.437±.077	.000±.000

of positions for Ranking Loss, 93.1% for Coverage Error, 89.2% for Average Precision, and 93.3% for Micro-F1; it is among the top three at every one of the 1,440 positions. The low-budget region is particularly relevant to feature selection because it tests whether useful variables are placed near the start of the ranking. Over budgets 1%–5%, the corresponding best rates are 98.9%, 100.0%, 87.8%, and 81.1%. Micro-F1 therefore exhibits somewhat greater early-budget variability than the loss metrics, but never falls outside the top three and attains the best budget-averaged value on every reported setting. Overall, the curves show that the table-level gains are distributed across the ranking rather than being produced by a few favorable budgets. FIND is not uniformly best at every budget: some baselines cross its trajectory at high budgets, while its average and low-budget performance remains strongest.

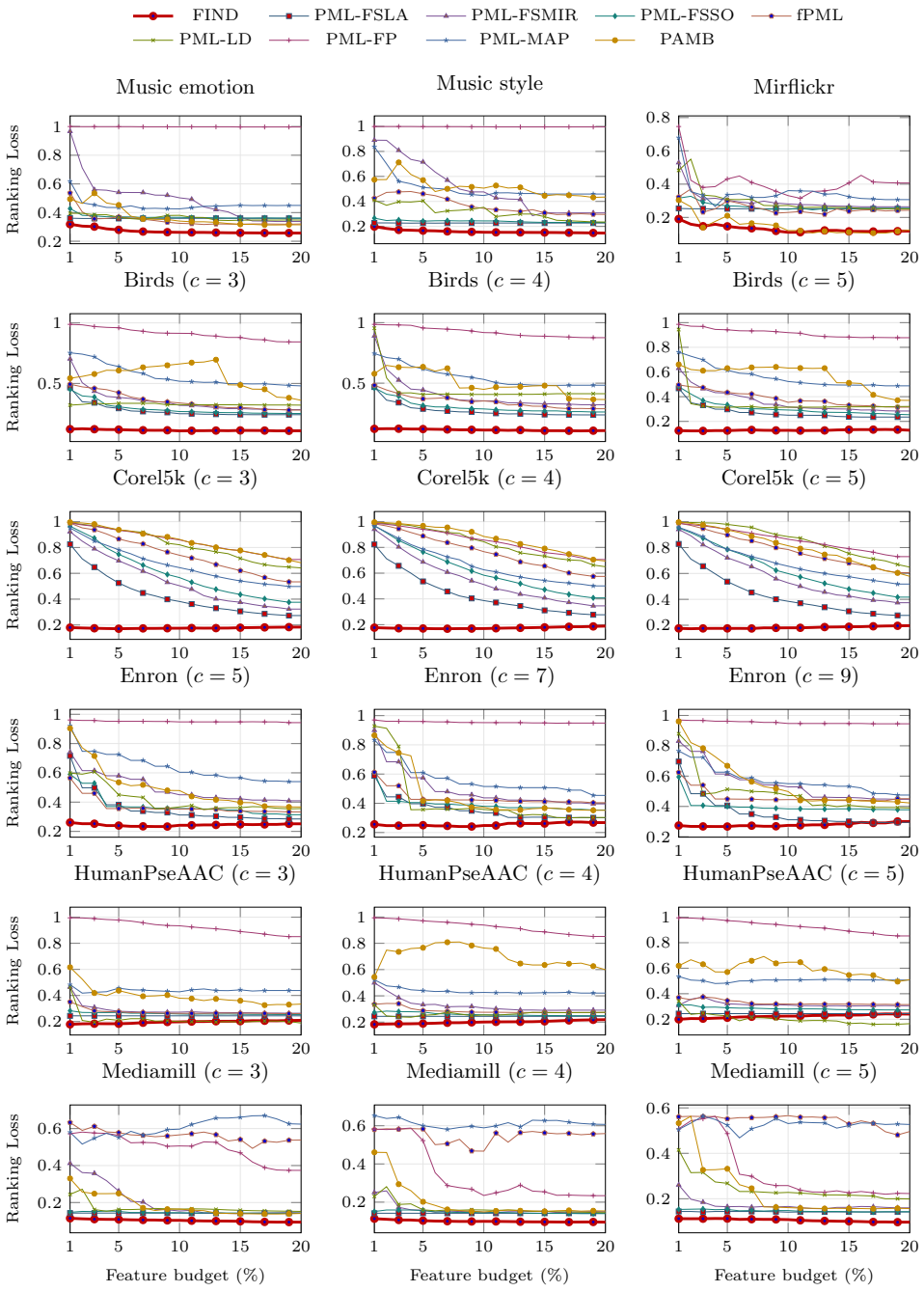


Figure 3: Ranking Loss across feature budgets from 1% to 20%. The first row contains the three real-world PML datasets; the remaining rows contain all 15 controlled settings, grouped by dataset family. Lower is better.

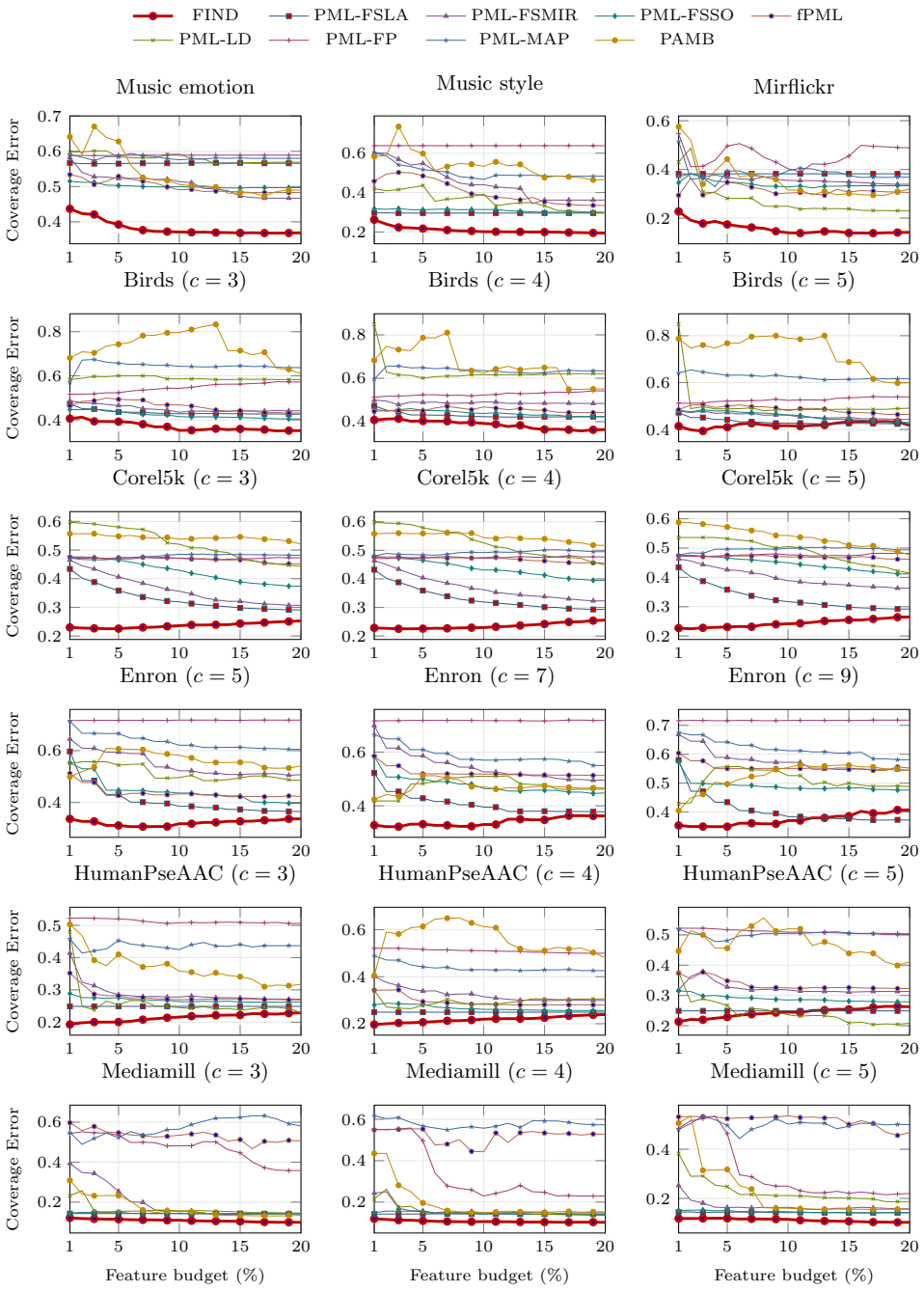


Figure 4: Coverage Error across feature budgets from 1% to 20% on the same 18 settings as Figure 3. Lower is better.

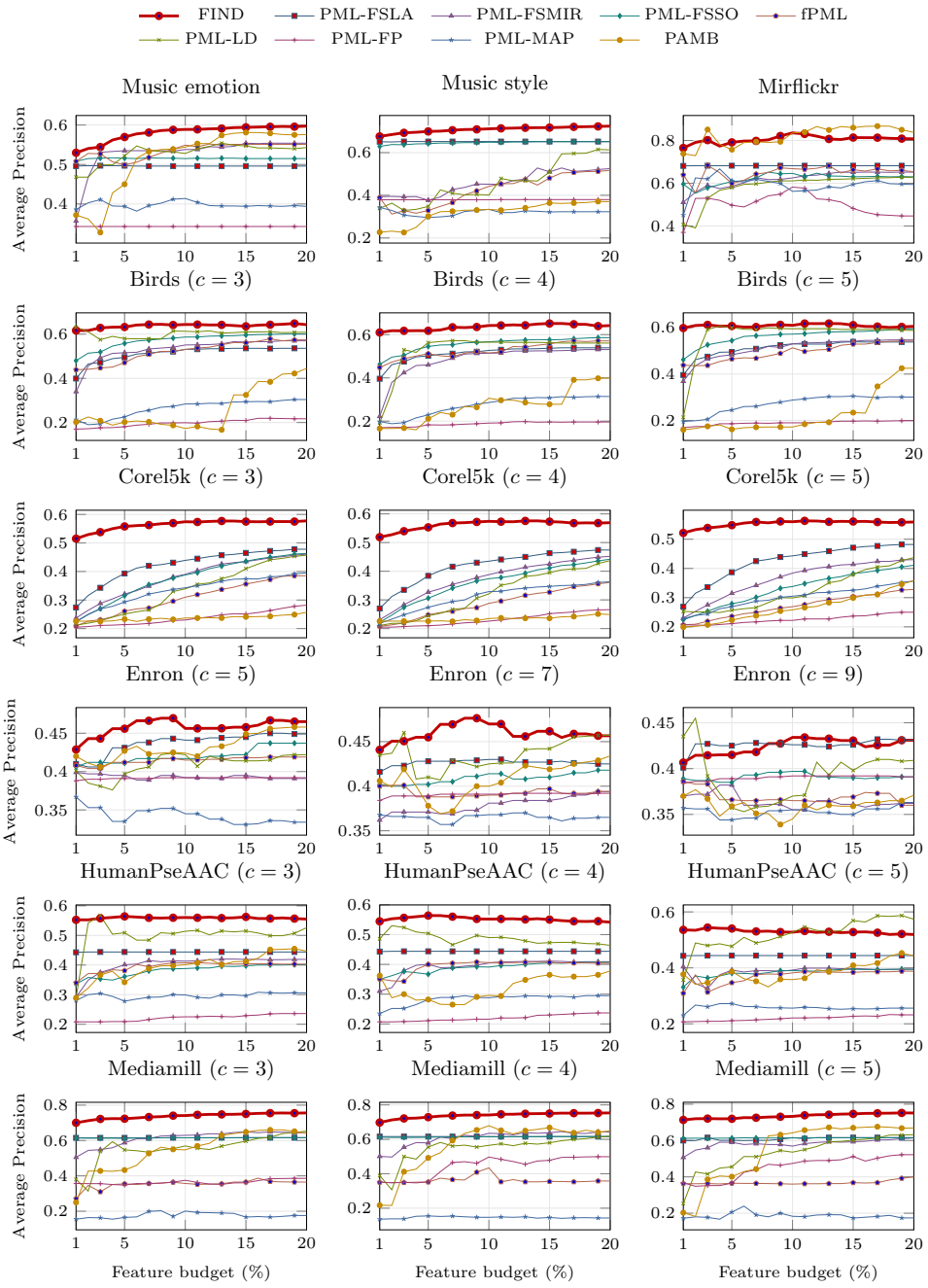


Figure 5: Average Precision across feature budgets from 1% to 20% on the same 18 settings as Figure 3. Higher is better.

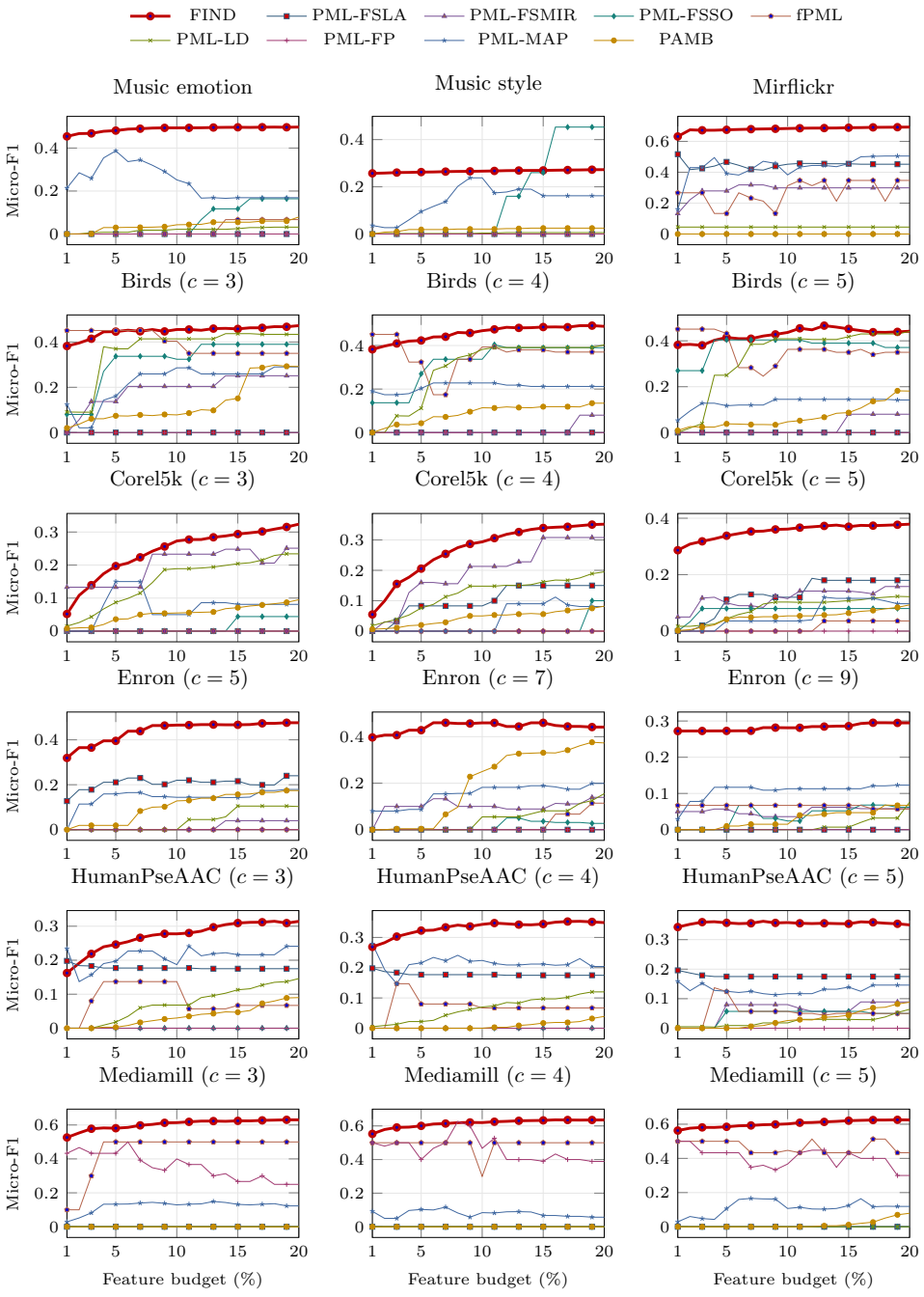


Figure 6: Micro-F1 across feature budgets from 1% to 20% on the same 18 settings as Figure 3. Higher is better.

D THEORETICAL RESULTS AND PROOFS

This appendix gives the projection and optimization details used to solve the ambiguity-aware objective, and collects the formal results supporting the set-deletion interpretation and the numerical certificates used by the algorithm.

D.1 ADMISSIBLE TARGETS AND THE SQUARED-DISTANCE LOSS

Proposition 3 (Convex representation and projection)**.** *For every nonempty candidate set s_i, the admissible target set satisfies*

$$\operatorname{conv}(\mathcal{Y}_i) = \left\{z \in [0,1]^q : 0 \le z \le s_i,\ \mathbf{1}^\top z \ge 1\right\} = \mathcal{Q}_i.$$

For any $p \in \mathbb{R}^q$, let $C_i = \{j : s_{ij} = 1\}$ and $u = \operatorname{clip}(p_{C_i}, 0, 1)$. The non-candidate coordinates of the projection are zero. If $\mathbf{1}^\top u \ge 1$, its candidate coordinates equal u; otherwise, they equal the Euclidean projection of the original candidate scores p_{C_i} onto $\Delta_i = \{z \ge 0 : \mathbf{1}^\top z = 1\}$.

Proof. After removing the fixed non-candidate coordinates, $\mathcal{Q}_i$ reduces to $P = \{z \in [0,1]^m : \mathbf{1}^\top z \ge 1\}$. Every vertex of P is binary and nonzero. Indeed, two fractional coordinates permit a feasible opposite perturbation and therefore cannot define a vertex. One fractional coordinate can also be perturbed if the sum constraint is inactive; if it is active, all other coordinates being binary forces the remaining coordinate to be either zero or one. Thus the vertices of P are exactly the nonzero binary vectors, which proves the convex-hull identity.

Projection onto the surrounding box is coordinate-wise clipping. If the clipped point satisfies the sum constraint, it is also the projection onto its subset P. Otherwise, the sum constraint must be active at the projection, reducing the problem to projection onto $\{z \ge 0 : \mathbf{1}^\top z = 1\}$. Its coordinates automatically do not exceed one, so the box upper bound introduces no additional constraint. □

Interpretation of the relaxation. The convex hull intentionally retains every label assignment compatible with the observed candidate set; in particular, any prediction $p_i \in \mathcal{Q}_i$ has zero sample loss. Thus, $\mathcal{Q}_i$ alone neither selects a unique latent target nor guarantees recovery of the unavailable ground truth. The inductive preference instead comes from fitting a single predictor jointly across samples under the capacity and sparsity regularizers in Eq. equation 4. Consequently, the deletion certificates characterize feature necessity relative to this explicit ambiguity-aware criterion, not identifiability of the true label assignment.

Proposition 4 (Loss regularity)**.** *The function $p \mapsto \frac{1}{2}\operatorname{dist}^2(p, \mathcal{Q}_i)$ is convex and continuously differentiable, with*

$$\nabla_p \frac{1}{2}\operatorname{dist}^2(p, \mathcal{Q}_i) = p - \Pi_{\mathcal{Q}_i}(p). \tag{30}$$

Its gradient is one-Lipschitz. Consequently, the objective F is proper, coercive, and γ-strongly convex, and every unconstrained or feature-deleted problem has a unique minimizer.

Proof. The squared distance to a nonempty closed convex set is the Moreau envelope of its indicator function. The standard envelope identity gives the stated gradient, and firm non-expansiveness of Euclidean projection gives its one-Lipschitz continuity. Convexity is preserved under the affine predictor. The $\ell_{2,1}$ term is convex, while $\frac{\gamma}{2}(\|W\|_F^2 + \|b\|_2^2)$ is γ-strongly convex and coercive. The sum therefore has the claimed properties; restriction to the closed linear constraint $W_{G,:} = 0$ preserves existence and uniqueness. □

D.2 ACCELERATED PROXIMAL OPTIMIZATION

Write equation 4 as $F(W, b) = H(W, b) + \lambda\|W\|_{2,1}$, where H is smooth. At an extrapolated iterate $(Z_W^{(t)}, Z_b^{(t)})$, the gradient of the loss is evaluated row-wise using equation 30. Given a step size α, let

$$V^{(t)} = Z_W^{(t)} - \alpha \nabla_W H(Z_W^{(t)}, Z_b^{(t)}).$$

The coefficient update is the row-wise group soft-thresholding operator

$$W_{k,:}^{(t+1)} = \left(1 - \frac{\alpha\lambda}{\|V_{k,:}^{(t)}\|_2}\right)_+ V_{k,:}^{(t)}, \tag{31}$$

while b is updated by an ordinary gradient step. A short power iteration estimates $\|[X, \mathbf{1}]\|_2^2/n$; denoting its 1.05-inflated value by $\widehat{L}$, the implementation uses $\alpha = 0.98/(\widehat{L} + \gamma)$. Adaptive restart suppresses oscillation. These choices control numerical convergence only: the certificates in Section 3.2 are based on the returned stationarity residual rather than on a presumed convergence rate.

The full and feature-deleted problems share this proximal solver. For a deleted set G, the rows $W_{G,:}$ are reset to zero after both the proximal and extrapolation steps. The intercept-only reference has no group penalty or design multiplication and is solved by the reduced b-gradient iteration with step size $0.98/(1 + \gamma)$, costing $O(nq)$ per iteration.

D.3 Deletion Characterization and Structural Properties

Proof of Theorem 1. If $D(G) \leq \epsilon$, let $\theta^\star_{-G}$ denote the constrained optimizer. By definition, $F(\theta^\star_{-G}) = F^\star + D(G) \leq F^\star + \epsilon$. It therefore belongs to $\mathcal{A}_\epsilon$ while avoiding G. Conversely, any $(W, b) \in \mathcal{A}_\epsilon$ with $W_{G,:} = 0$ is feasible for the constrained problem, so $F^\star_{-G} \leq F(W, b) \leq F^\star + \epsilon$, which gives $D(G) \leq \epsilon$. Taking the logical negation gives the equivalent form. □

Corollary 5 (Individual core and collective emergence)**.** *The individually necessary features satisfy*

$$\mathcal{N}_\epsilon = \{k : D(\{k\}) > \epsilon\} = \bigcap_{(W,b)\in\mathcal{A}_\epsilon} \operatorname{supp}(W).$$

Moreover, if $D(\{k\}) \leq \epsilon$ for every $k \in G$ but $D(G) > \epsilon$, every member of G is individually avoidable while no near-optimal model can avoid all of G.

Proof. Apply Theorem 1 first with $G = \{k\}$ and then with the entire set G. □

Proposition 6 (Monotonicity, normalization, and scale invariance)**.** *For $G_1 \subseteq G_2$,*

$$0 = D(\varnothing) \leq D(G_1) \leq D(G_2) \leq D(\{1, \ldots, d\}) = B.$$

Hence $\rho(G) = D(G)/B \in [0, 1]$ when $B > 0$. In addition, $\epsilon_1 \leq \epsilon_2$ implies $\mathcal{N}_{\epsilon_2} \subseteq \mathcal{N}_{\epsilon_1}$. Finally, replacing F by $aF + c$ for $a > 0$ leaves $\rho(G)$ and the decision $\rho(G) > \eta$ unchanged.

Proof. The feasible set under $W_{G_2,:} = 0$ is contained in that under $W_{G_1,:} = 0$, proving monotonicity of the constrained optimum. Empty deletion recovers $F^\star$, whereas deleting all features leaves precisely the intercept-only problem and therefore costs $F_0^\star - F^\star = B$. The normalized bounds follow immediately. Increasing ϵ can only change inequalities of the form $D(\{k\}) > \epsilon$ from true to false. Under the positive affine transformation, both $D(G)$ and B are multiplied by a, while the additive constant cancels, so their ratio and threshold decision are invariant. □

D.4 Safe Numerical Certification

For completeness, the minimum-norm subgradient in equation 16 is available in closed form. Let $A_{k,:} = \nabla_{W_{k,:}} H(\widehat{W}, \widehat{b})$. Its kth coefficient row is

$$U_{k,:} = \begin{cases} A_{k,:} + \lambda \widehat{W}_{k,:}/\|\widehat{W}_{k,:}\|_2, & \|\widehat{W}_{k,:}\|_2 > 0, \\ (1 - \lambda/\|A_{k,:}\|_2)_+ A_{k,:}, & \|\widehat{W}_{k,:}\|_2 = 0, \end{cases} \tag{32}$$

with the second line defined as zero when $A_{k,:} = 0$. Thus $r^2 = \sum_k \|U_{k,:}\|_2^2 + \|\nabla_b H(\widehat{W}, \widehat{b})\|_2^2$.

For a constrained re-optimization with $W_{G,:} = 0$, stationarity is evaluated on the free coordinates rather than in the ambient parameter space. Specifically, let $A^G_{k,:} = \nabla_{W_{k,:}} H(\widehat{W}^G, \widehat{b}^G)$ and define $U^G_{k,:}$ by the two cases in equation 32 for $k \notin G$. The constrained residual and its certified gap are

$$r_G^2 = \sum_{k\notin G} \|U^G_{k,:}\|_2^2 + \|\nabla_b H(\widehat{W}^G, \widehat{b}^G)\|_2^2, \qquad g_G = \frac{r_G^2}{2\gamma}. \tag{33}$$

The fixed rows $k \in G$ are excluded from the residual norm; their generally nonzero ambient gradients are therefore not mistaken for lack of convergence of the constrained problem.

Proof of Theorem 2. Let $\widehat{\theta} = (\widehat{W}, \widehat{b})$ and choose the minimum-norm $s \in \partial F(\widehat{\theta})$, so $\|s\|_2 = r$. Strong convexity gives

$$F(\theta^\star) \geq F(\widehat{\theta}) + \langle s, \theta^\star - \widehat{\theta} \rangle + \frac{\gamma}{2}\|\theta^\star - \widehat{\theta}\|_2^2.$$

Maximizing the resulting quadratic upper bound over $\|\theta^\star - \widehat{\theta}\|_2$ yields $\widehat{F} - F^\star \leq r^2/(2\gamma) = g$. Strong monotonicity of ∂F also gives $\|\widehat{\theta} - \theta^\star\|_2 \leq r/\gamma = R$.

For any feasible θ satisfying $W_{G,:} = 0$, strong convexity at the optimum gives $F(\theta) - F^\star \geq \frac{\gamma}{2}\|\theta - \theta^\star\|_2^2$. Minimizing over this constraint gives

$$D(G) \geq \frac{\gamma}{2}\|W^\star_{G,:}\|_F^2 \geq \frac{\gamma}{2}(\|\widehat{W}_{G,:}\|_F - R)_+^2.$$

The zeroed model $(\widetilde{W}^G, \widehat{b})$ is feasible, while $F^\star \geq \widehat{F} - g$, proving the upper bound.

The two certified objective intervals imply

$$\max\{0, \widehat{F}_0 - g_0 - \widehat{F}\} \leq B \leq \max\{0, \widehat{F}_0 - \widehat{F} + g\},$$

which proves equation 19. Comparing enclosing intervals proves the two safe decisions. Finally, $F^\star_{-G} \in [\widehat{F}_G - g_G, \widehat{F}_G]$ and $F^\star \in [\widehat{F} - g, \widehat{F}]$; subtracting their endpoints and using $D(G) \geq 0$ gives a second valid interval. Intersecting it with the initial interval preserves containment of $D(G)$ and gives equation 23. □

For the numerical implementation of these certificates, the full and feature-deleted proximal-gradient fits use an initial stopping tolerance of 2×10^{-5} and are initially limited to 140 and 90 iterations, respectively. The specialized intercept-only solver is limited to 3,000 iterations and stops directly by its computable gap criterion. Certification requires the computable gap to satisfy $g/\epsilon \leq 0.05$ (and $g_G/\epsilon \leq 0.05$ for a deleted fit). If this condition is not met, the solver performs at most three continuation rounds of up to 1,000 iterations, reducing the optimization tolerance by a factor of ten per round down to 10^{-9}. These stopping rules control numerical accuracy but do not replace the interval decision: a deletion is declared removable or non-removable only when its certified interval lies entirely on one side of $[\epsilon_{\mathrm{L}}, \epsilon_{\mathrm{U}}]$; otherwise it remains unresolved.

D.5 Correctness of the Nested Collective Search

Proposition 7 (Nested-prefix search)**.** *Let $r_1, \ldots, r_m$ be individually removable features and $G_j = \{r_1, \ldots, r_j\}$. Then the predicate $D(G_j) > \epsilon$ is monotone in j. Given exact tests, exponential bracketing followed by binary search finds the smallest non-removable prefix in $O(\log m)$ group tests after a crossing is bracketed. With interval tests, every reported removable or non-removable endpoint is valid; unresolved tests yield a bracket rather than an unsupported minimality claim.*

Proof. The sets are nested, so Proposition 6 gives $D(G_j) \leq D(G_{j+1})$. The exact predicate is therefore false before a single transition and true afterwards, which is the standard invariant maintained by exponential bracketing and binary search. Under interval tests, equation 22 certifies each declared endpoint. An unresolved midpoint establishes neither side of the exact predicate, so retaining the current certified bracket is the only sound update and preserves the stated guarantee. □

D.6 Complete Certification and Discovery Procedure

Algorithm 1 summarizes the complete certification and discovery procedure described in Sections 3.2–3.3. The optimization subroutine follows Appendix D.2.

Algorithm 1 FIND: Feature-necessity certification and collective discovery

Require: Data (X, S); parameters λ, γ, η; maximum collective-search prefix $K_{\max}$
Ensure: Singleton partition $(\widehat{\mathcal{N}}, \widehat{\mathcal{R}}, \widehat{\mathcal{U}})$, predictive ranking π, and optional collective certificate

Solve equation 4 to obtain $(\widehat{W}, \widehat{b})$ and $\widehat{F}$
Compute the stationarity gap g and radius R
Solve the intercept-only problem to obtain $\widehat{F}_0$ and gap g_0
Compute $[\epsilon_{\mathrm{L}}, \epsilon_{\mathrm{U}}]$ using equation 19
function CERTIFY(G)
    Compute $[\underline{D}(G), \overline{D}(G)]$
    if $\underline{D}(G) > \epsilon_{\mathrm{U}}$ **then**
        return NONREMOVABLE
    else if $\overline{D}(G) \leq \epsilon_{\mathrm{L}}$ **then**
        return REMOVABLE
    end if
    Re-optimize equation 8 subject to $W_{G,:} = 0$
    Compute the intersected interval $[\underline{D}_{\mathrm{final}}(G), \overline{D}_{\mathrm{final}}(G)]$ using equation 23
    if $\underline{D}_{\mathrm{final}}(G) > \epsilon_{\mathrm{U}}$ **then**
        return NONREMOVABLE
    else if $\overline{D}_{\mathrm{final}}(G) \leq \epsilon_{\mathrm{L}}$ **then**
        return REMOVABLE
    else
        return UNRESOLVED
    end if
end function
Initialize $\widehat{\mathcal{N}}, \widehat{\mathcal{R}}, \widehat{\mathcal{U}} \leftarrow \varnothing$
for $k = 1, \ldots, d$ **do**
    $z_k \leftarrow$ CERTIFY($\{k\}$)
    Assign k to $\widehat{\mathcal{N}}$, $\widehat{\mathcal{R}}$, or $\widehat{\mathcal{U}}$ according to z_k
end for
Construct π by placing $\widehat{\mathcal{N}}$ first and sorting each block by decreasing $\|\widehat{W}_{k,:}\|_2$ (stable index ties)
if $\widehat{\mathcal{N}} \neq \varnothing$ **or** $\widehat{\mathcal{U}} \neq \varnothing$ **then**
    return singleton partition, π, and no group claim
end if
Order $\widehat{\mathcal{R}} = (r_1, \ldots, r_M)$ by decreasing $\|\widehat{W}_{r_j,:}\|_2$
Define $G_j = \{r_1, \ldots, r_j\}$, $M' \leftarrow \min(M, K_{\max})$; $j_L \leftarrow 1$, $j_U \leftarrow \varnothing$, $j \leftarrow 2$
if $M' < 2$ **then**
    return singleton partition, π, and no collective set
end if
while $j \leq M'$ **do**
    $z \leftarrow$ CERTIFY(G_j)
    if $z =$ REMOVABLE **then**
        $j_L \leftarrow j$
    else if $z =$ NONREMOVABLE **then**
        $j_U \leftarrow j$; **break**
    end if
    if $j = M'$ **then**
        break
    end if
    $j \leftarrow \min(2j, M')$
end while
if $j_U = \varnothing$ **then**
    return singleton partition, π, and no certified collective set within the search range
end if
while $j_U - j_L > 1$ **do**
    $j \leftarrow \lfloor (j_L + j_U)/2 \rfloor$; $z \leftarrow$ CERTIFY(G_j)
    if $z =$ REMOVABLE **then**
        $j_L \leftarrow j$
    else if $z =$ NONREMOVABLE **then**
        $j_U \leftarrow j$
    else
        break
    end if
end while
return singleton partition, π, certified bracket (j_L, j_U), and collective set G_{j_U}