CO-EVOLUTIONARY PROBABILISTIC RECOVERY FOR MULTI-DIMENSIONAL CLASSIFICATION WITH MISSING LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-dimensional classification (MDC) requires assigning labels to each instance across multiple class dimensions simultaneously, but fully labeling all dimensions is often costly or even impractical. Thus, a setting with missing labels more closely approximates the reality of MDC. Despite its practical relevance, learning from incomplete labels has yet to be investigated in MDC. Existing label completion strategies developed for other learning paradigms cannot be directly applied to MDC, as they typically rely on predefined label correlations or independent recovery assumptions, while overlooking the fact that missing labels may simultaneously obscure the dependencies among multiple class dimensions. Moreover, inaccurate completion can further distort the dependency information available to the learner and introduce unreliable supervision for downstream classifier learning. To address these challenges, we propose CSER, a competitive co-evolutionary framework for probabilistic label recovery. CSER first constructs complementary feature-only and dependency-conditioned class-probability representations from observed labels, and then co-evolves two interacting populations of recovery strategies and pseudo-missing challenges. The challenge population searches for difficult recovery targets that expose weaknesses of current strategies, while the recovery population learns confidence-aware, dimension-wise fusion preferences. A robust strategy is finally selected against accumulated historical challenges and used to recover labels for genuinely missing entries, with multiple imputation propagating the recovered supervision to downstream MDC learning. Experiments on 17 benchmark datasets demonstrate that CSER significantly outperforms MDC baselines equipped with label completion strategies across various label-missing scenarios.

1 INTRODUCTION

Multi-dimensional classification (MDC) is a supervised learning paradigm in which each instance is associated with multiple class variables drawn from heterogeneous class spaces Read et al. (2014); Jia & Zhang (2024). Compared with conventional single-output classification, MDC provides a more complete description of an object by predicting several related but semantically different class values simultaneously. These output dimensions are not isolated: their labels may exhibit co-occurrence, shared feature patterns, and conditional relationships that can provide useful cross-dimensional information for prediction. Exploiting such dependencies while preserving the semantics of each class dimension is a central characteristic of MDC Zhang et al. (2026). While such a formulation has been successfully applied to text and sentiment analysis Dinan et al. (2020), visual analysis Lian et al. (2020), and recommendation He & Gu (2025), its multi-dimensional nature also makes complete annotation costly or impractical in practice. Experts may annotate only case-relevant dimensions, different annotation sources may cover different subsets of labels, and some class dimensions may be costly or even infeasible to annotate. As a result, in many real-world scenarios, the training label matrix can be only partially observed: an instance may retain valid labels on several dimensions while missing labels on others. Despite its practical relevance, learning from incomplete labels has yet to be systematically investigated in MDC. We refer to this setting as

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.

incomplete-label MDC, in which the learner must infer unobserved class assignments from partially observed labels while preserving informative dependencies across class dimensions.

Existing MDC methods have developed various strategies to exploit dependencies among class dimensions. Representative approaches include explicit modeling dependency by Bayesian networks and classifier chains Gil-Begue et al. (2021); Read et al. (2021); Jia & Zhang (2022a), implicit dependency modeling by feature augmentation Jia & Zhang (2020a; 2022b) and label-space encoding Jia & Zhang (2021; 2023), as well as evolutionary optimization of chain structures and feature subsets Zhang et al. (2025a). Despite their methodological differences, these approaches are generally built on the assumption that the class vector of each training instance is fully observed. This assumption prevents their direct application to MDC with missing labels, where only partial class labels are available for learning cross-dimensional dependencies and training predictive models.

A common solution to label missingness is to first estimate the unknown labels and then train a predictor on the completed label matrix. Matrix-completion and correlation-based approaches infer missing entries by exploiting low-rank, local, or correlation structures Cheng & Tan (2025); Liu et al. (2025b). More recent methods further improve missing-label recovery through prototype modeling, soft-label estimation, or explicit missing-label probability modeling Liu et al. (2024a); Zhao et al. (2026); Fu et al. (2026). Although these techniques provide useful guidance, directly filling missing labels across all dimensions may be risky. First, dimensional heterogeneity causes the reliability of predictions based on inputs and observable labels to vary substantially across dimensions and instances. Second, exploiting dependencies among labels may improve recovery performance, but inaccurate auxiliary predictions may also introduce errors. Third, the ground-truth values of genuinely missing labels are unobservable, making it impossible to directly assess the reliability of candidate recovery strategies at these positions. Therefore, the central challenge in incomplete MDC is not merely how to estimate a missing label, but how to learn a recovery strategy that remains reliable across heterogeneous output dimensions and diverse missingness configurations.

Motivated by this observation, we recast label recovery in incomplete-label MDC as a **C**o-evolution between mi**S**singness chall**E**nge discovery and **R**ecovery-strategy learning, and propose **CSER** to instantiate this paradigm. Instead of directly assigning pseudo-labels to unknown entries under a fixed missingness pattern, CSER constructs pseudo-missing tasks by temporarily hiding selected observed labels, thereby providing verifiable guidance for learning how to leverage cross-dimensional dependencies. It then constructs two interacting populations around these tasks: an adversarial missingness challenge population that evolves informative pseudo-missing configurations from observed labels, whereas a probabilistic recovery strategy population that evolves how feature-based and cross-dimensional predictive evidence should be integrated for label recovery. Through their competitive interaction, the two populations are then optimized in opposite directions: recovery strategies are encouraged to reduce recovery error across evolving challenges, while missingness configurations are encouraged to expose weaknesses of the current recovery population. This reciprocal process enables the missing-label environment and the recovery behavior to adapt to each other without iteratively injecting predicted labels into training. After co-evolution, a robust recovery strategy is selected against the accumulated historical challenges and used to infer probabilistic labels for the genuinely missing entries, providing completed supervision for downstream MDC learning. The main contributions of this work are summarized as follows:

- **Verifiable challenge-driven recovery for MDC with missing labels:** CSER constructs pseudo-missing tasks from observed labels to provide verifiable recovery feedback, and further evolves these tasks as a challenge population to expose weaknesses of candidate recovery strategies, so that the two tasks are cooperatively optimized.
- **Confidence-aware probabilistic recovery:** CSER constructs complementary feature-only and dependency-conditioned class-probability representations from observed labels, and learns dimension-wise recovery preferences to integrate them according to instance-specific confidence. Historical archive-based paired evaluation selects the final recovery strategy, while multiple imputation propagates the recovered probabilistic supervision to downstream MDC learning.
- **Extensive evaluation under incomplete-label settings:** Experiments on 17 benchmark datasets with 30%, 50%, and 70% label missingness show consistent improvements over four representative MDC methods combined with two label-completion strategies, with statistically significant gains in almost all comparisons.

2 Related Work

Multi-Dimensional Classification. Multi-dimensional classification (MDC) aims to predict a vector of discrete class variables drawn from multiple heterogeneous class spaces. A central research topic is how to exploit dependencies among these class dimensions to improve predictive performance. Under fully supervised settings, Bayesian-network classifiers and classifier-chain methods explicitly model inter-dimensional dependencies Zaragoza et al. (2011); Gil-Begue et al. (2021). Decomposition and stacking approaches use predictions from one subproblem to assist others Jia & Zhang (2022a; 2020b), while feature augmentation and label-space transformation methods capture dependency information implicitly Jia & Zhang (2020a; 2021; 2023); Liu et al. (2024b). Recent studies further investigate dimension-specific feature learning Huang et al. (2024); Zhang et al. (2025b), reduce unnecessary dependency assumptions Huang et al. (2025), or optimize dependency structures and feature subsets through evolutionary search Zhang et al. (2025a). Despite their different modeling strategies, these methods generally assume that the training class vector is fully observed.

Semi-supervised MDC incorporates unlabeled training instances to reduce the dependence on fully labeled data Huang et al. (2023); Xu & Li (2023). However, it assumes that each training instance is either fully labeled or fully unlabeled, which differs from the incomplete-label setting considered in this work, where labels are missing at the individual class-dimension level. Therefore, existing semi-supervised MDC methods cannot directly recover and utilize partially observed label vectors.

Missing-Label Learning in Related Paradigms. Missing-label learning has been extensively studied in related paradigms, especially multi-label learning, where missing entries are recovered through low-rank structures Yu et al. (2014), label correlations Cheng & Tan (2025); Rastogi & Mortaza (2021), concept-based structures Liu et al. (2025b), soft-label estimation Zhao et al. (2026), or probabilistic modeling Fu et al. (2026). Related efforts in multi-view multi-label learning and application-specific settings further exploit representation learning Liu et al. (2025a), prototype modeling Liu et al. (2024a), or auxiliary semantic information to infer incomplete annotations Ma et al. (2024). Although these studies provide valuable insights for incomplete supervision, they assume output structures and dependency patterns different from MDC. Their recovery mechanisms can be adapted as external completion modules, but they do not address the MDC-specific problem of recovering missing labels while preserving dependencies among heterogeneous class dimensions.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{X} = \mathbb{R}^d$ denote the feature space and $\mathcal{Y} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_q$ the MDC output space, where $\mathcal{C}_j = \{1, \ldots, K_j\}$ is the class space of the j-th dimension. Given a feature matrix $\mathbf{X} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$ and a label matrix $\mathbf{Y} = [y_1, \ldots, y_n]^\top \in \mathbb{Z}^{n \times q}$, where $y_i = (y_{i1}, \ldots, y_{iq})^\top \in \mathcal{Y}$ and $y_{ij} \in \mathcal{C}_j$, only a subset of entries in $\mathbf{Y}$ is observed. We define the observation mask $\mathbf{M} \in \{0, 1\}^{n \times q}$ as $M_{ij} = \mathbf{1}\{y_{ij} \text{ is observed}\}$. The observed and genuinely missing label entries are defined as

$$\Omega^o = \{(i, j) \mid M_{ij} = 1\}, \qquad \Omega^m = \{(i, j) \mid M_{ij} = 0\}. \tag{1}$$

For the j-th class dimension, we denote the indices of training instances with observed labels as $\mathcal{I}_j^o = \{i \mid (i, j) \in \Omega^o\}$. The learner has access to $\mathbf{X}$, $\mathbf{M}$, and the observed labels $\{y_{ij} \mid (i, j) \in \Omega^o\}$. CSER constructs pseudo-missing challenges $H \subseteq \Omega^o$ by selecting a portion of observed entries as recovery targets. Their labels are used to score recovery strategies. Such a challenge specifies evaluation positions in a fixed probability representation. It does not remove labels from the estimator training sets or change $\mathbf{M}$. Notably, labels in Ω^m are unavailable throughout learning.

3.2 Overview of CSER

Recovering missing labels in MDC is challenging because the reliability of feature-based and cross-dimensional evidence varies across dimensions and instances. As illustrated in Fig. 1, CSER addresses this problem in three stages. First, it constructs two fixed class-probability representations from observed training labels, capturing feature-only and dependency-conditioned evidence, respectively. Second, CSER competitively co-evolves a recovery-strategy population and a pseudo-missing

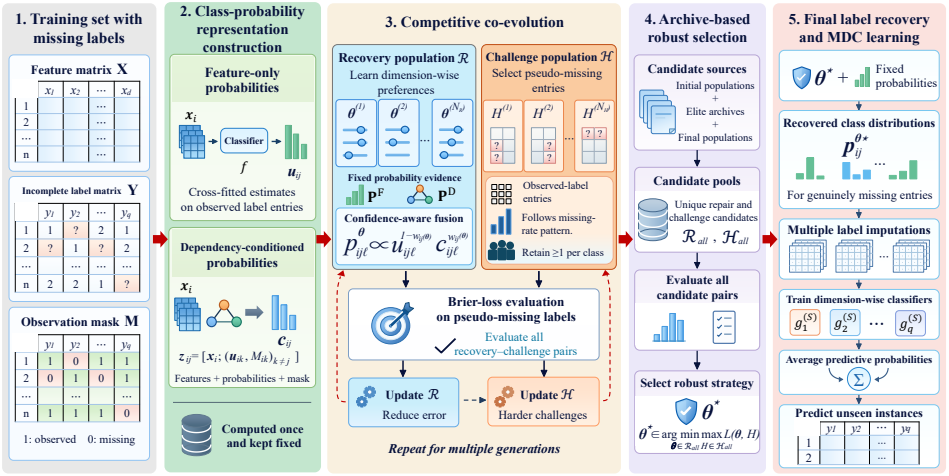


Figure 1: Overview of the proposed CSER framework.

challenge population: recovery strategies learn to integrate the two probability representations, while challenges search for difficult observed-label subsets that expose their recovery weaknesses. Finally, the selected recovery strategy generates the training label matrix through multiple imputations by sampling from the recovered class probability distribution. These completed supervisions are used to train the MDC predictor, which evaluates unseen test instances without accessing the genuinely missing labels.

3.3 Class-Probability Representation Construction

Reliable label recovery requires informative class-probability representations from both instance features and cross-dimensional dependencies. CSER therefore constructs complementary feature-only and dependency-conditioned representations, $\mathbf{P}^{\mathrm{F}} = \{\boldsymbol{u}_{ij}\}_{i=1,j=1}^{n,q}$ and $\mathbf{P}^{\mathrm{D}} = \{\boldsymbol{c}_{ij}\}_{i=1,j=1}^{n,q}$. Each $\boldsymbol{u}_{ij}$ and $\boldsymbol{c}_{ij}$ is a K_j-dimensional class-probability vector over $\mathcal{C}_j$. The two representations are generated by probabilistic classifiers trained only on observed labels and can be applied to both observed and genuinely missing entries. For an observed entry, the prediction is cross-fitted and does not use that entry's own label; for a genuinely missing entry, it is the average of the predictions from all fold-specific models.

The feature-only representation estimates each dimension from $\boldsymbol{x}_i$. To avoid evaluating observed targets with models fitted to their own labels, CSER uses held-out predictions at observed positions and ensemble predictions at missing positions:

$$\boldsymbol{u}_{ij} = \frac{1}{B}\sum_{b=1}^{B}\begin{cases} f^{\mathrm{F}}_{jb,-\phi^{\mathrm{F}}_{jb}(i)}(\boldsymbol{x}_i), & (i,j)\in\Omega^o, \\ \dfrac{1}{F^{\mathrm{F}}_{jb}}\displaystyle\sum_{k=1}^{F^{\mathrm{F}}_{jb}} f^{\mathrm{F}}_{jb,-k}(\boldsymbol{x}_i), & (i,j)\in\Omega^m. \end{cases} \tag{2}$$

Here B is the number of repetitions, F^{F}_{jb} is the fold count, $\phi^{\mathrm{F}}_{jb}(i)$ identifies the held-out fold, and $f^{\mathrm{F}}_{jb,-k}$ is trained on observed labels outside fold k.

To capture cross-dimensional information while retaining uncertainty in other dimensions' predictions, the dependency-conditioned representation uses

$$\boldsymbol{z}_{ij} = [\boldsymbol{x}_i; (\boldsymbol{u}_{ik}, M_{ik})_{k\neq j}]\,. \tag{3}$$

Here, the auxiliary information from other dimensions is not the observed class labels, but rather feature-based probability representations and observation indicators. M_{ik} distinguish observed labels from the MDC sample with genuinely missing label. The vector $\boldsymbol{c}_{ij}$ follows the same two-case averaging rule as Eq. 2, replacing $\boldsymbol{x}_i$ and the feature-only models with $\boldsymbol{z}_{ij}$ and dependency-conditioned models, respectively. Appendix A.1 specifies the implementation.

Both representations remain fixed during co-evolution and recovery. The evolutionary process only learns how to exploit and combine these probabilistic representations for missing-label estimation.

3.4 Label Recovery Strategy Representation

Incomplete labels may obscure the true dependencies among label dimensions, making the reliability of constructed probabilistic representations vary across dimensions. To this end, CSER parameterizes a recovery strategy by a dimension-wise weighting vector

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q) \in [0,1]^q, \tag{4}$$

where θ_j specifies the preference for the dependency-conditioned representation when recovering the j-th label dimension.

Given a recovery strategy $\boldsymbol{\theta}$, CSER adaptively adjusts the contribution of each probability representation according to its confidence. Specifically, the confidence of each probability vector is measured by the standard normalized entropy-based confidence score r_j, where lower entropy indicates higher confidence. Then, the confidence-aware fusion weight and probability are defined as

$$w_{ij}(\boldsymbol{\theta}) = \frac{\theta_j r_j(\boldsymbol{c}_{ij})}{\theta_j r_j(\boldsymbol{c}_{ij}) + (1-\theta_j) r_j(\boldsymbol{u}_{ij})}, \quad p^{\boldsymbol{\theta}}_{ij\ell} = \frac{u_{ij\ell}^{1-w_{ij}(\boldsymbol{\theta})} c_{ij\ell}^{w_{ij}(\boldsymbol{\theta})}}{\sum_{a=1}^{K_j} u_{ija}^{1-w_{ij}(\boldsymbol{\theta})} c_{ija}^{w_{ij}(\boldsymbol{\theta})}}. \tag{5}$$

Here, $w_{ij}(\boldsymbol{\theta})$ controls the adaptive contribution of the dependency-conditioned probability representation for sample i and label dimension j. For any entry $(i,j) \in \Omega^o \cup \Omega^m$, the fused distribution

$$\boldsymbol{p}^{\boldsymbol{\theta}}_{ij} = (p^{\boldsymbol{\theta}}_{ij\ell})_{\ell=1}^{K_j} \tag{6}$$

gives the recovered class probabilities, where $p^{\boldsymbol{\theta}}_{ij\ell}$ denotes the probability that sample i belongs to class ℓ in the j-th label dimension under strategy $\boldsymbol{\theta}$. The resulting distributions are used to evaluate recovery quality on pseudo-missing entries and to complete genuinely missing labels. The numerically stabilized implementation of the entropy-weighted probability fusion is detailed in Appendix A.2.

Accordingly, each individual in the recovery population $\mathcal{R}$ represents a label recovery strategy: $\mathcal{R} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N_R)}\}$, where N_R is the population size. The fitness of each strategy is evaluated according to the recovery performance obtained from the above probability fusion mechanism on pseudo-missing entries. The optimized strategy is then applied to genuinely missing entries for final label recovery.

3.5 Pseudo-Missing Challenge Representation

Although probabilistic representations provide complementary evidence for label recovery, a recovery strategy optimized on fixed recovery targets may overfit specific cases and degrade on other difficult entries. To learn more robust recovery strategies, CSER introduces pseudo-missing challenges that expose candidate strategies to varying subsets of observed-label recovery targets.

Let $\mathcal{I}^o_{j\ell} = \{i \in \mathcal{I}^o_j : y_{ij} = \ell\}$ be the observed instance indices of class ℓ in dimension j, with $n_{j\ell} = |\mathcal{I}^o_{j\ell}|$, and let $\rho_j = n^{-1}\sum_i (1 - M_{ij})$ be the empirical missing rate of dimension j. A pseudo-missing challenge is represented by class-wise selection sets $H_{j\ell} \subseteq \mathcal{I}^o_{j\ell}$, $j = 1, \ldots, q$, $\ell = 1, \ldots, K_j$, specifying the observed entries selected as recovery targets. They jointly determine the challenge entry set:

$$H = \bigcup_{j,\ell} \{(i,j) : i \in H_{j\ell}\} \subseteq \Omega^o. \tag{7}$$

Before co-evolution, class-wise quotas $h_{j\ell}$ are fixed according to the empirical missing rates and observed class support. Every challenge satisfies

$$|H_{j\ell}| = h_{j\ell}, \qquad \sum_{\ell=1}^{K_j} h_{j\ell} \approx \rho_j |\mathcal{I}^o_j|, \qquad 0 \le h_{j\ell} \le (n_{j\ell} - 1)_+, \tag{8}$$

where $(a)_+ = \max(a, 0)$. The target count $\rho_j|\mathcal{I}_j^o|$ is adjusted for rounding and support constraints, leaving at least one observed entry of each represented class outside the challenge. Exact quota allocation and boundary cases are detailed in Appendix A.3.

CSER maintains a fixed-size challenge population $\mathcal{H} = \{H^{(1)}, \ldots, H^{(N_H)}\}$. To evaluate recovery strategies under diverse missingness scenarios, challenges are initialized by class-wise sampling and evolved under the fixed quotas in Eq. 8. Therefore, CSER searches for recovery strategies that remain robust under diverse difficult missing-label scenarios rather than fitting a specific missingness pattern.

3.6 Competitive Co-Evolution for Challenge-Guided Label Recovery

To avoid optimizing against a single missing-label pattern, CSER maintains diverse populations of recovery strategies and challenges, allowing it to accumulate complementary experience cases. It couples the recovery and challenge populations through competitive co-evolution: recovery strategies seek accurate recovery under the current challenges, whereas challenges search for feasible pseudo-missing subsets that reveal weaknesses by the current recovery population. Before evolution, each recovery strategy is initialized by independently sampling its dimension-wise weights θ_j from $[0, 1]$. Each challenge is initialized by sampling the class-wise selection sets $H_{j\ell}$ from $\mathcal{I}_{j\ell}^o$ without replacement according to the prescribed quotas.

The two populations interact through a common recovery loss. For a recovery strategy $\boldsymbol{\theta}$, the fixed representations $\boldsymbol{u}_{ij}$ and $\boldsymbol{c}_{ij}$ are fused according to Eq. 5 and 6 to produce $\boldsymbol{p}_{ij}^{\boldsymbol{\theta}}$. For a challenge H with per-dimension quota $h_j = \sum_{\ell=1}^{K_j} |H_{j\ell}|$ and $\mathcal{J} = \{j : h_j > 0\}$, we define the entry-wise Brier loss $e_{ij}(\boldsymbol{\theta})$, and the recovery loss $L(\boldsymbol{\theta}, H)$, taken first within and then across dimensions in $\mathcal{J}$:

$$
\begin{aligned}
e_{ij}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_{\ell=1}^{K_j} \left(p_{ij\ell}^{\boldsymbol{\theta}} - \mathbf{1}\{y_{ij} = \ell\}\right)^2, \\
L(\boldsymbol{\theta}, H) &= \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \frac{1}{h_j} \sum_{i \in \bigcup_{\ell=1}^{K_j} H_{j\ell}} e_{ij}(\boldsymbol{\theta}).
\end{aligned}
\tag{9}
$$

where $p_{ij\ell}^{\boldsymbol{\theta}}$ is the ℓ-th component of $\boldsymbol{p}_{ij}^{\boldsymbol{\theta}} = (p_{ij\ell}^{\boldsymbol{\theta}})_{\ell=1}^{K_j}$, and $y_{ij} \in \mathcal{C}_j$ is the true label at (i, j), which is observed since $(i, j) \in H \subseteq \Omega^o$ but treated as missing only for the purpose of this evaluation. This normalized Brier loss evaluates the full predictive distribution, thereby accounting for probabilistic confidence beyond the predicted class label.

Let $\mathcal{R} = \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N_R)}\}$ and $\mathcal{H} = \{H^{(1)}, \ldots, H^{(N_H)}\}$ denote the current recovery and challenge populations. We define the fitness F_R of a recovery strategy $\boldsymbol{\theta}$ as its mean recovery loss L against every challenge in $\mathcal{H}$, and the fitness F_H of a challenge H as its mean recovery loss against every strategy in $\mathcal{R}$:

$$
F_R(\boldsymbol{\theta}; \mathcal{H}) = \frac{1}{|\mathcal{H}|} \sum_{H \in \mathcal{H}} L(\boldsymbol{\theta}, H), \quad F_H(H; \mathcal{R}) \quad = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{\theta} \in \mathcal{R}} L(\boldsymbol{\theta}, H),
\tag{10}
$$

Here, individuals for label recovery are selected to minimize F_R, as it measures recovery error, whereas individuals to specify challenge maximize F_H by seeking pseudo-missing configurations that pose greater recovery difficulty in the current recovery population. Accordingly, the recovery population is driven toward strategies effective across current difficult cases, while the challenge population continually searches for configurations that expose the remaining weaknesses of recovery strategies. These two conflicting goals are repeatedly optimized through competitive co-evolution, and CSER continuously updates these two populations across successive generations.

In generation t, the two populations are updated sequentially using the population-averaged fitness in Eq. 10. Repair offspring are first generated from $\mathcal{R}^{t-1}$ and, together with their parents, evaluated against $\mathcal{H}^{t-1}$; the N_R individuals with the lowest mean losses form $\mathcal{R}^t$. Challenge offspring are then generated from $\mathcal{H}^{t-1}$ and, together with their parents, evaluated against the updated $\mathcal{R}^t$; the N_H individuals with the highest mean losses form $\mathcal{H}^t$. The challenge variation operators and sequential population-update procedure are detailed in Appendices A.3 and A.4. After both updates, their individuals are re-evaluated against the updated opponent population, and the best candidates under the corresponding objectives are stored in the archives.

3.7 Probabilistic Label Recovery and Inductive MDC Learning

Let $\mathcal{R}_{\text{all}} = \mathcal{R}^0 \cup \mathcal{R}^T \cup \mathcal{A}_R$ and $\mathcal{H}_{\text{all}} = \mathcal{H}^0 \cup \mathcal{H}^T \cup \mathcal{A}_H$ collect the distinct recovery strategies and challenges from the initial populations, the populations after T generations, and the elite archives. CSER selects the strategy that minimizes the worst-case recovery loss over $\mathcal{H}_{\text{all}}$:

$$\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}\in\mathcal{R}_{\text{all}}} \max_{H\in\mathcal{H}_{\text{all}}} L(\boldsymbol{\theta}, H), \tag{11}$$

where ties are broken by mean loss over $\mathcal{H}_{\text{all}}$ and then candidate order.

The selected strategy $\boldsymbol{\theta}^\star$ combines $\mathbf{P}^{\text{F}}$ and $\mathbf{P}^{\text{D}}$ at genuinely missing positions through Eq. 6. Nevertheless, directly filling every missing label with a single fixed value is risky: because recovery reliability is heterogeneous across dimensions and instances, a hard imputation would treat an uncertain guess as ground truth, and cross-dimensional dependencies could then propagate its error into other dimensions during downstream training. Instead, CSER propagates the probabilistic recovery information to downstream learning through S completed label matrices $\widetilde{\mathbf{Y}}^{(s)} = [\widetilde{y}_{ij}^{(s)}]$. For each $\ell = 1, \ldots, K_j$,

$$\Pr\left(\widetilde{y}_{ij}^{(s)} = \ell\right) = \begin{cases} \mathbf{1}\{y_{ij} = \ell\}, & (i,j) \in \Omega^o, \\ p_{ij\ell}^{\boldsymbol{\theta}^\star}, & (i,j) \in \Omega^m, \end{cases} \qquad s = 1, \ldots, S. \tag{12}$$

Thus, observed labels are preserved, while each missing label is independently sampled from its recovered class distribution across imputations.

For each imputation s and dimension j, we train a probabilistic classifier $g_j^{(s)}$ on $\{(\boldsymbol{x}_i, \widetilde{y}_{ij}^{(s)})\}_{i=1}^n$ using features alone. Unlike the feature-only estimators used for evidence construction, these classifiers are fitted to all n instances with completed labels and require no outer cross-fitting. At inference time, the MDC predictor averages the class probabilities produced by the S imputations and selects, for each dimension, the class with the largest average. Notably, imputed labels are used only for downstream learning, while the probability representations used for recovery remain fixed. Further details on archive-based robust selection and multiple imputation are provided in Appendix A.5. In this way, CSER turns incomplete training supervision into completed supervision for inductive MDC prediction on unseen instances.

4 Experiments

4.1 Benchmark Datasets and Missing-Label Settings

We evaluate all methods on 17 benchmark MDC datasets with diverse sample sizes and output structures. The datasets range from several hundred to more than ten thousand examples and include both low- and high-dimensional class spaces. Their per-dimension class cardinalities also vary considerably. This diversity enables evaluation under different data scales, numbers of class dimensions, and label-space complexities.

In this experiment, we divided each dataset into training, validation, and test sets in a $7:1:2$ ratio. Incomplete supervision is simulated only on the training label matrix, while the test labels remain complete for evaluation. For each dataset, we generate missing-label settings with nominal rates of 30%, 50%, and 70%. At a missing rate ρ, approximately ρ of the training labels are hidden and the remaining labels are retained as observed supervision. Five missing-label realizations are used for each rate, corresponding to different seeds. All compared methods use the same training/test partition and the same missing-label mask for each combination of dataset, missing rate, and repetition, enabling faired comparisons across methods.

Table 1: Basic information of the 17 experimental datasets.

	Flare1	Jura	Enb	Song	WQplants	WQanimals	WaterQuality	BeLaE	Voice	Rf1	Thyroid	Pain	CoIL2000	TIC2000	Fera	Adult
#Examples	323	359	768	785	1060	1060	1060	1930	3136	8987	9172	9734	9822	9822	14052	18419
#Dimensions	3	2	2	3	7	7	14	5	2	8	7	10	5	3	5	4
#Labels/Dim.	3,4,2	4,5	2,4	3	4	4	4	5	4,2	4,4,3,4,4,3,4,3	5,5,3,2,4,4,3	2,5,4,2,2,5,2,5,2,2	6,10,10,4,2	6,4,2	6	7,7,5,2
#Features	10x	9n	6n	98n	16n	16n	16n	1n,44x	19n	64n	7n,22x	136n	81x	83x	136n	5n,5x

4.2 Compared Approaches

We compare CSER with four state-of-the-art MDC methods: DCOM Huang et al. (2025), PIST Huang et al. (2024), LEAD Zhang et al. (2025b), and ECCO Zhang et al. (2025a). Among them, ECCO is also an evolutionary MDC baseline because it jointly optimizes classifier-chain structure and dimension-specific feature subsets. DCOM reduces reliance on explicit class dependency modeling through dependency decoupling, PIST learns pairwise dimension-specific features, and LEAD exploits label correlations to guide label-specific feature learning. Since all existing MDC methods are originally designed for complete training labels, the experimental workspace provides two preprocessing strategies for missing labels: prototype-based imputation Zhao et al. (2026); Liu et al. (2024a) and correlation-based imputation Cheng & Tan (2025); Rastogi & Mortaza (2021). Detailed implementations of the two label-imputation strategies used for baseline adaptation are provided in Appendix B. The eight baselines formed in this way use exactly the same train/test partitions, missing-label masks and random seeds.

4.3 Evaluation Metrics

All methods first recover the missing training labels and then train an MDC predictor for unseen test instances. Given a test set $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{p}$ with q label dimensions, let $\hat{y}_{ij}$ and y_{ij} denote the predicted and ground-truth labels, respectively, and define $r^{(i)} = \sum_{j=1}^{q} [\![y_{ij} = \hat{y}_{ij}]\!]$. Following standard MDC evaluation practice Jia & Zhang (2024), we evaluate performance using Hamming Score (HS), Exact Match (EM), and Sub-Exact Match (SEM):

$$\mathrm{HS}_{\mathcal{S}} = \frac{1}{p}\sum_{i=1}^{p}\frac{r^{(i)}}{q}, \quad \mathrm{EM}_{\mathcal{S}} = \frac{1}{p}\sum_{i=1}^{p}[\![r^{(i)} = q]\!], \quad \mathrm{SEM}_{\mathcal{S}} = \frac{1}{p}\sum_{i=1}^{p}[\![r^{(i)} \geq q-1]\!]. \tag{13}$$

4.4 Implementation Details

All methods are evaluated under identical data splits and missing-label realizations. PIST uses the validation set for model selection following its original procedure, whereas DCOM, LEAD, ECCO, and CSER do not access validation data during training or model selection. Competing MDC methods retain the classifiers, hyperparameters, and training procedures of their original implementations. In CSER, all probability representations are constructed from observed training labels using out-of-fold prediction and remain fixed during co-evolution.

CSER uses a linear-kernel LIBSVM classifier for downstream MDC learning. Both populations are evolved for 50 generations with population size 50. The sensitivity to the evolution budget is analyzed in Appendix E. Complete hyperparameter settings, evolutionary operators, and computational complexity are provided in Appendices A.6 and A.7.

The Python-based methods are executed on an NVIDIA GeForce RTX 4090 GPU, whereas the MATLAB-based methods are run on an Intel Core i7-14700K workstation with 20 parallel workers.

5 Results and Analysis

The Holm-adjusted Wilcoxon results in Table 2 demonstrate that the advantage of CSER remains statistically reliable after correcting for multiple comparisons. Among the 72 pairwise tests across three missing rates, three metrics, and eight competing methods, CSER obtains 70 significant wins, two ties, and no significant loss. Together with the consistently best average ranks on HS, EM, SEM in Tables S-I– S-VIII in Appendix C, these results indicate that the improvement is broadly maintained across datasets, metrics, and missing-label settings.

More importantly, as shown in Table 3, the advantage of CSER becomes more pronounced as label missingness increases. The large performance gaps are also consistent with the substantial performance degradation of competing MDC methods equipped with popular completion strategies under severe label missingness. Specifically, incorrectly completed labels are passed directly to downstream MDC learners and may be further propagated through cross-dimensional dependency modeling. In contrast, CSER learns recovery strategies through verifiable pseudo-missing challenges and performs confidence-aware probabilistic recovery rather than committing to a fixed completion rule.

Table 2: Holm-adjusted Wilcoxon signed-rank test results of CSER against eight compared methods.

Missing Rate	Metric	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
30%	HS	**win**[2.14E-04]	**win**[3.01E-03]	**win**[6.10E-04]	**win**[7.32E-04]	**win**[7.32E-04]	**win**[1.34E-03]	**win**[1.22E-04]	**win**[5.49E-04]
	EM	**win**[1.53E-03]	**win**[3.78E-03]	**win**[5.46E-03]	**win**[3.78E-03]	**win**[1.28E-03]	**win**[3.78E-03]	**win**[3.66E-04]	**win**[1.28E-03]
	SEM	**win**[1.71E-03]	**tie**[5.17E-02]	**win**[8.24E-03]	**tie**[5.17E-02]	**win**[8.24E-03]	**win**[1.15E-02]	**win**[1.22E-04]	**win**[2.38E-03]
50%	HS	**win**[1.22E-04]	**win**[3.20E-04]	**win**[1.22E-04]	**win**[3.20E-04]	**win**[1.22E-04]	**win**[2.44E-04]	**win**[1.22E-04]	**win**[3.20E-04]
	EM	**win**[1.83E-04]	**win**[9.16E-04]	**win**[1.83E-04]	**win**[1.16E-03]	**win**[1.22E-04]	**win**[4.88E-04]	**win**[1.22E-04]	**win**[4.88E-04]
	SEM	**win**[3.05E-04]	**win**[5.37E-03]	**win**[2.14E-04]	**win**[1.28E-03]	**win**[2.14E-04]	**win**[6.10E-04]	**win**[1.22E-04]	**win**[1.28E-03]
70%	HS	**win**[1.22E-04]	**win**[1.22E-04]	**win**[1.22E-04]	**win**[1.22E-04]	**win**[1.22E-04]	**win**[1.22E-04]	**win**[1.22E-04]	**win**[1.22E-04]
	EM	**win**[1.22E-04]	**win**[1.83E-04]	**win**[1.22E-04]	**win**[1.83E-04]	**win**[1.22E-04]	**win**[1.83E-04]	**win**[1.22E-04]	**win**[1.22E-04]
	SEM	**win**[3.05E-04]	**win**[1.10E-03]	**win**[2.44E-04]	**win**[1.10E-03]	**win**[2.44E-04]	**win**[1.10E-03]	**win**[2.44E-04]	**win**[6.10E-04]

Table 3: Test Hamming Score (HS) under 70% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.949±0.000**	0.936±0.006	0.937±0.002	0.945±0.006	0.948±0.002	0.848±0.059	0.947±0.003	0.905±0.070	0.943±0.005
Jura	**0.618±0.024**	0.478±0.045	0.438±0.061	0.369±0.080	0.416±0.106	0.458±0.026	0.432±0.059	0.470±0.050	0.404±0.066
Enb	**0.758±0.044**	0.681±0.016	0.521±0.099	0.685±0.022	0.407±0.018	0.656±0.014	0.431±0.018	0.665±0.012	0.423±0.043
Song	**0.757±0.016**	0.700±0.072	0.697±0.011	0.743±0.018	0.686±0.000	0.678±0.016	0.686±0.001	0.682±0.031	0.686±0.001
WQplants	**0.706±0.002**	0.641±0.086	0.704±0.000	0.660±0.058	0.703±0.000	0.530±0.018	0.703±0.001	0.548±0.018	0.704±0.000
WQanimals	**0.662±0.002**	0.606±0.054	0.660±0.000	0.585±0.043	0.660±0.000	0.404±0.025	0.660±0.001	0.407±0.027	0.660±0.000
WaterQuality	**0.691±0.002**	0.633±0.029	0.690±0.000	0.649±0.019	0.690±0.000	0.466±0.023	0.690±0.000	0.470±0.030	0.683±0.015
BeLaE	**0.399±0.015**	0.349±0.007	0.353±0.004	0.287±0.011	0.351±0.007	0.343±0.009	0.353±0.004	0.357±0.010	0.351±0.005
Voice	**0.897±0.021**	0.673±0.009	0.372±0.037	0.606±0.008	0.293±0.020	0.684±0.013	0.399±0.026	0.660±0.052	0.277±0.000
Rf1	**0.750±0.020**	0.590±0.017	0.515±0.021	0.587±0.013	0.510±0.026	0.529±0.040	0.532±0.037	0.549±0.040	0.475±0.041
Thyroid	**0.968±0.001**	0.957±0.000	0.957±0.000	0.957±0.000	0.957±0.000	0.889±0.006	0.958±0.000	0.946±0.015	0.957±0.000
Pain	**0.959±0.000**	0.958±0.000	0.958±0.000	0.957±0.000	0.958±0.000	0.833±0.015	0.958±0.000	0.891±0.029	0.939±0.043
ColL2000	**0.898±0.004**	0.799±0.005	0.659±0.011	0.768±0.007	0.581±0.000	0.752±0.004	0.633±0.002	0.669±0.036	0.582±0.000
TIC2000	**0.897±0.004**	0.808±0.003	0.767±0.013	0.811±0.004	0.687±0.001	0.738±0.011	0.713±0.002	0.707±0.026	0.686±0.000
Fera	**0.661±0.003**	0.532±0.000	0.532±0.000	0.538±0.009	0.532±0.000	0.449±0.006	0.541±0.000	0.439±0.031	0.532±0.000
Adult	**0.703±0.003**	0.642±0.007	0.538±0.001	0.648±0.004	0.536±0.001	0.631±0.006	0.535±0.001	0.514±0.035	0.490±0.012
Default	**0.666±0.002**	0.490±0.003	0.603±0.000	0.514±0.005	0.603±0.000	0.400±0.000	0.455±0.000	0.603±0.001	0.603±0.001
Avg. Rank	**1.000**	4.471	4.706	4.706	5.706	6.941	4.882	6.206	6.382

This design allows the recovery process to adapt to heterogeneous evidence and difficult missing-label configurations before the recovered supervision is propagated to downstream learning. The resulting improvements on unseen test instances further indicate that the recovered supervision is effective for inductive MDC prediction. Additional analyses of challenge evolution and recovery behavior under evolving challenges are provided in Appendix F.

CSER also shows a small standard deviation across five repeated runs with practical computational cost (Appendix C, Tables S-IX–S-XI). Its consistent gains on HS, EM, and SEM demonstrate improvements in both dimension-wise and joint MDC prediction on unseen instances.

Table 4: Ablation summary on Wilcoxon signed-rank test under all dataset–setting comparisons.

CSER versus	HS (win/tie/loss)	EM (win/tie/loss)	SEM (win/tie/loss)
CSER-H	3/0/0	3/0/0	3/0/0
CSER-E	3/0/0	3/0/0	3/0/0

Ablation study. We evaluate two variants: **CSER-H** replaces probabilistic recovery with deterministic completion, while **CSER-E** removes competitive co-evolution and uses a fixed recovery rule. In Table 4, their performance degradation confirms the complementary roles of probabilistic recovery and co-evolutionary strategy optimization. Detailed settings and results are provided in Appendix D.

6 Conclusion

We present the first systematic study of MDC with missing labels and propose CSER, a competitive co-evolutionary framework for verifiable label recovery. CSER constructs complementary feature-only and dependency-conditioned class-probability representations and combines them through dimension-wise recovery preferences and instance-specific confidence. Verifiable pseudo-missing challenges co-evolve with recovery strategies to expose difficult recovery conditions, after which the selected probabilistic strategy provides recovered supervision for inductive MDC prediction. Experiments on 17 benchmark datasets across three missing-label rates demonstrate consistent advantages over representative MDC baselines equipped with label-completion strategies.

AI USE STATEMENT

In this study, we used generative AI tools to improve the clarity, fluency, and readability of the English writing. In addition, we used generative AI tools to implement part of the code functionality. We did not use generative AI tools for the conception or design of the research, nor for data analysis, interpretation of results, or derivation of conclusions. The remaining required disclosure tasks are also not applicable to this study. We have reviewed all AI-assisted work. Specifically, the authors reviewed and revised the LLM-assisted English writing to ensure that the original scientific meaning was preserved, and verified and tested the LLM-assisted code functionality to ensure its correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, sensitive personal data, or experiments that could raise ethical concerns. The datasets used in this study are publicly available, and no privacy or security issues are implicated. Our study focuses purely on methodological and computational aspects. Accordingly, we do not anticipate any direct ethical or societal risks arising from this research. We have no conflicts of interest to declare, and the research was conducted in compliance with applicable legal and research integrity requirements.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our results. The descriptions of the proposed models and algorithms are included in the main text, while additional implementation details, hyperparameter settings, and training procedures are provided in the appendix and supplementary material. Information about datasets and data preprocessing steps is clearly documented. To further facilitate reproducibility, we provide an anonymous compressed file containing the main source code, experiment scripts, and configuration files in the supplementary materials.

REFERENCES

Ziwei Cheng and Zhenhua Tan. Multi-label learning for label-specific features using correlation information with missing label. *Expert Systems with Applications*, 269:126491, 2025.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pp. 314–331, Virtual Event, 2020. ACL.

Xiaozhen Fu, Deyu Li, Yanhui Zhai, Yuhua Qian, and Suge Wang. A robust multi-label learning method based on missing label probability modeling. *Information Sciences*, pp. 123459, 2026.

Santiago Gil-Begue, Concha Bielza, and Pedro Larrañaga. Multi-dimensional bayesian network classifiers: A survey. *Artificial Intelligence Review*, 54(1):519–559, 2021.

Zhengqiang He and Yuanyuan Gu. Research on recommendation framework for personalized push of cultural and creative products in tourist cities based on multi-level computational user profiling. *Journal of Combinatorial Mathematics and Combinatorial Computing*, 127:4837–4856, 2025.

Teng Huang, Bin-Bin Jia, and Min-Ling Zhang. Progressive label propagation for semi-supervised multi-dimensional classification. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pp. 3821–3829, Macao, SAR, China, 2023. IJCAI.

Teng Huang, Bin-Bin Jia, and Min-Ling Zhang. Deep multi-dimensional classification with pairwise dimension-specific features. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 4183–4191, Jeju Island, South Korea, 2024. IJCAI.

Teng Huang, Bin-Bin Jia, and Min-Ling Zhang. Towards escaping from class dependency modeling for multi-dimensional classification. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 25356–25371, Vancouver, BC, Canada, 2025. PMLR.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via kNN feature augmentation. *Pattern Recognition*, 106:107423, 2020a.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via stacked dependency exploitation. *Science China Information Sciences*, 63(12):222102, 2020b.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via sparse label encoding. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4917–4926, Virtual Event, 2021. PMLR.

Bin-Bin Jia and Min-Ling Zhang. Decomposition-based classifier chains for multi-dimensional classification. *IEEE Transactions on Artificial Intelligence*, 3(2):176–191, 2022a.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via selective feature augmentation. *Machine Intelligence Research*, 19(1):38–51, 2022b.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification via decomposed label encoding. *IEEE Transactions on Knowledge and Data Engineering*, 35(2):1844–1856, 2023.

Bin-Bin Jia and Min-Ling Zhang. Multi-dimensional classification: Paradigm, algorithms and beyond. *Vicinagearth*, 1(1):3, 2024.

Zheng Lian, Ya Li, Jian-Hua Tao, Jian Huang, and Ming-Yue Niu. Expression analysis based on face regions in real-world conditions. *International Journal of Automation and Computing*, 17 (1):96–107, 2020.

Chengliang Liu, Gehui Xu, Jie Wen, Yabo Liu, Chao Huang, and Yong Xu. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32253–32267. PMLR, 2024a.

Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4940–4956, 2025a.

Xinyuan Liu, Jihua Zhu, Zhiqiang Tian, and Zhongyu Li. Multi-dimensional classification via class space fusion and comprehensive label correlations. *Information Fusion*, 111:102521, 2024b.

Zhiming Liu, Jinhai Li, Xiao Zhang, and Xizhao Wang. Multi-level information fusion for missing multi-label learning based on stochastic concept clustering. *Information Fusion*, 115:102775, 2025b. doi: 10.1016/j.inffus.2024.102775.

Leilei Ma, Hongxing Xie, Lei Wang, Yanping Fu, Dengdi Sun, and Haifeng Zhao. Text-region matching for multi-label image recognition with missing labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6133–6142. ACM, 2024. doi: 10.1145/3664647.3680815.

Jonathan Ortigosa-Hernández, Juan Diego Rodríguez, Leandro Alzate, Manuel Lucania, Inaki Inza, and Jose A Lozano. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92:98–115, 2012.

Reshma Rastogi and Sayed Mortaza. Multi-label classification with missing labels using label correlation and robust structural learning. *Knowledge-Based Systems*, 229:107336, 2021.

Jesse Read, Luca Martino, and David Luengo. Efficient monte carlo methods for multi-dimensional learning with classifier chains. *Pattern Recognition*, 47(3):1535–1546, 2014.

Jesse Read, Bernhard Pfahringer, Geoffrey Holmes, and Eibe Frank. Classifier chains: A review and perspectives. *Journal of Artificial Intelligence Research*, 70:683–718, 2021.

Zelin Xu and Peipei Li. Decomposition and combination-based classifier chains for semi-supervised multi-dimensional classification. In *Proceedings of the IEEE International Conference on Knowledge Graph*, pp. 184–191, Shanghai, China, 2023. IEEE.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit S. Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 593–601. PMLR, 2014.

Julio H. Zaragoza, L. Enrique Sucar, Eduardo F. Morales, Concha Bielza, and Pedro Larrañaga. Bayesian chain classifiers for multidimensional classification. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 2192–2197, Barcelona, Catalonia, Spain, 2011.

Yu-Yang Zhang, Bin-Bin Jia, and Min-Ling Zhang. Evolutionary classifier chain for multi-dimensional classification. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pp. 22641–22649, Philadelphia, PA, USA, 2025a. AAAI Press.

Yu-Yang Zhang, Bin-Bin Jia, and Min-Ling Zhang. Evolutionary dimension-specific feature selection for multi-dimensional classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4290–4302, 2026.

Zan Zhang, Jialin Zhou, Jialu Yao, Lin Liu, Jiuyong Li, Lei Li, and Xindong Wu. Learning label-specific features for multi-dimensional classification. *Pattern Recognition*, pp. 112365, 2025b.

Dawei Zhao, Hong Li, Yixiang Lu, De Zhu, and Qingwei Gao. Anchor-based graph embedding and soft label learning for multi-label classification with missing label. *Expert Systems with Applications*, 296:129019, 2026.

A CSER: Detailed Algorithms and Implementation

This appendix specifies the numerical and boundary cases of the main-text construction. Throughout, $\boldsymbol{\theta}$ and H denote label recovery and challenge individuals, $\mathcal{R}$ and $\mathcal{H}$ their populations, and N_R, N_H their prescribed sizes. Populations are indexed collections that may contain duplicates, while sizes and fitness averages count every individual. Only the final candidate pools are deduplicated. We reserve t for evolutionary generations and use $\tau \in \{\mathrm{F}, \mathrm{D}\}$ for the two probability-estimation stages.

A.1 Cross-Validated Class-Probability Estimation

The inputs are the training features $\mathbf{X}$, the original observation mask $\mathbf{M}$, and $\mathbf{Y}_{\Omega^o}$. Raw class labels are encoded within each dimension as $\mathcal{C}_j = \{1, \ldots, K_j\}$ and decoded after prediction.

Let $k_j = \sum_{\ell=1}^{K_j} \mathbf{1}\{n_{j\ell} > 0\}$ denote the number of observed classes. To distinguish intermediate predictions from the final stored representations, write $\bar{\boldsymbol{u}}_{ij}$ and $\bar{\boldsymbol{c}}_{ij}$ for the feature-stage and dependency-stage outputs before the final fallback rules. In particular, the dependency inputs are

$$\boldsymbol{z}_{ij} = [\boldsymbol{x}_i; (\bar{\boldsymbol{u}}_{ik}, M_{ik})_{k \neq j}]\,. \tag{14}$$

This makes explicit the construction-time meaning of the auxiliary probabilities in Eq. 3. The final $\boldsymbol{u}_{ij}, \boldsymbol{c}_{ij}$ retain their main-text meaning as the stored vectors used for fusion. They can differ from the stage outputs after stabilization or a single-observed-class override. $\boldsymbol{z}_{ij}$ is not rebuilt after these operations.

For each label dimension, CSER estimates feature-only and dependency-conditioned class probabilities using repeated cross-validation. For an observed entry, the probability vector is obtained from a model that excludes the fold containing that entry, whereas for a genuinely missing entry, predictions from all fold-specific models are averaged. Let $\boldsymbol{v}_{ij}^{\mathrm{F}} = \boldsymbol{x}_i$ and $\boldsymbol{v}_{ij}^{\mathrm{D}} = \boldsymbol{z}_{ij}$. For stage $\tau \in \{\mathrm{F}, \mathrm{D}\}$, repetition b, and dimension j, define

$$\boldsymbol{b}_{ijb}^{\tau} = \begin{cases} f_{jb,-\phi_{jb}^{\tau}(i)}^{\tau}(\boldsymbol{v}_{ij}^{\tau}), & (i,j) \in \Omega^o, \\ \dfrac{1}{F_{jb}^{\tau}} \displaystyle\sum_{k=1}^{F_{jb}^{\tau}} f_{jb,-k}^{\tau}(\boldsymbol{v}_{ij}^{\tau}), & (i,j) \in \Omega^m, \end{cases} \tag{15}$$

where F_{jb}^{τ} is the number of nonempty folds and $\phi_{jb}^{\tau}(i)$ denotes the fold containing observed instance i. Each $f_{jb,-k}^{\tau}$ is trained only on the observed target labels outside fold k.

To ensure numerically valid probability vectors, let $\Pi_\delta(\cdot)$ denote a numerical stabilization mapping that replaces invalid or near-zero components as necessary and renormalizes the resulting vector to a valid probability distribution, with $\delta = 10^{-12}$. The repetition-averaged estimate is then

$$\bar{\boldsymbol{p}}_{ij}^{\tau} = \Pi_\delta\left(\frac{1}{B}\sum_{b=1}^{B} \Pi_\delta(\boldsymbol{b}_{ijb}^{\tau})\right). \tag{16}$$

Accordingly, $\bar{\boldsymbol{p}}_{ij}^{\mathrm{F}} = \bar{\boldsymbol{u}}_{ij}$ and $\bar{\boldsymbol{p}}_{ij}^{\mathrm{D}} = \bar{\boldsymbol{c}}_{ij}$ for the feature-only and dependency-conditioned stages, respectively.

Folds are constructed independently for each dimension, repetition, and estimation stage using class-aware partitioning, with the fold count adapted to the available class support. Fold-specific outputs are aligned to the common class ordering $\mathcal{C}_j$. For dimensions with only one observed class, the corresponding representation reduces to a point-mass distribution, when dependency-conditioned estimation is unavailable, CSER falls back to the feature-only representation and marks the dependency evidence as unavailable. These rules ensure that every training entry receives a valid probability representation while only available dependency-conditioned evidence participates in the subsequent fusion. Algorithm A1 summarizes the two-stage cross-validated construction and finalization of the fixed class-probability representations.

Regarding the scope of held-out exclusion, the held-out prediction statement in Section 3.3 applies to each fold model's training targets. The procedure is not nested cross-fitting: all second-stage folds

Algorithm A1 Construction of Fixed Class-Probability Representations

Require: $\mathbf{X}, \mathbf{Y}_{\Omega^o}, \mathbf{M}$, class vocabularies, repetitions B, nominal folds F
Ensure: Fixed $\mathbf{P}^{\mathrm{F}}, \mathbf{P}^{\mathrm{D}}$, availability flags, and confidence/log-probability caches

Validate observed support and encode labels in each dimension.
Fit feature preprocessing on training features and apply it once.
for $\tau \in \{\mathrm{F}, \mathrm{D}\}$, in this order **do**
  if $\tau = \mathrm{D}$ **then**
    Construct $\boldsymbol{z}_{ij}$ from $\bar{\boldsymbol{u}}_{ik}$ and the original mask using Eq. 14.
  end if
  for $j = 1, \ldots, q$ **do**
    Use inputs $\boldsymbol{x}_i$ for $\tau = \mathrm{F}$ and $\boldsymbol{z}_{ij}$ for $\tau = \mathrm{D}$.
    for $b = 1, \ldots, B$ **do**
      Construct nonempty folds using the rule above.
      Fit each fold model outside its held-out fold, including the empty-training and single-class cases.
      Predict held-out observed rows and genuinely missing rows, and align class columns.
      Form $\boldsymbol{b}^{\tau}_{ijb}$ according to Eq. 15.
    end for
    If all repetitions succeed, average and stabilize their outputs using Eq. 16 to obtain $\bar{\boldsymbol{p}}^{\tau}_{ij}$.
    Abort on a feature-stage failure, and mark the dimension's dependency-conditioned output unavailable if any dependency repetition fails.
  end for
end for
for $j = 1, \ldots, q$ **do**
  if $k_j = 1$ **then**
    Replace both stage outputs by the point mass on the sole observed class, using all observed support in dimension j.
  end if
  Stabilize the feature-stage output to obtain $\boldsymbol{u}_{ij}$.
  Test the availability of the dependency-conditioned output after the override and record the corresponding availability flag.
  Set $\boldsymbol{c}_{ij}$ to the stabilized dependency-conditioned output when available, and to $\boldsymbol{u}_{ij}$ otherwise.
  Cache confidences, log probabilities, and the single-class flag.
end for
return Retain $\boldsymbol{z}_{ij}$'s original construction without recomputation.

reuse the same first-stage outputs, whose construction may involve labels from other dimensions of second-stage held-out rows. The dependency input excludes the target dimension's own feature-only vector and contains no raw auxiliary labels. Moreover, the $k_j = 1$ override uses all observed labels in that dimension and is therefore an exception to strict held-out prediction of the final stored representation. These rules do not imply independence of the entire construction from every observed label. Labels in Ω^m are never used, and pseudo-missing challenges do not alter estimator training. After finalization, co-evolution, archive-based selection, and genuine label recovery reuse the fixed representations without further estimator calls or changes to $\mathbf{M}$ or $\boldsymbol{z}_{ij}$.

A.2 Entropy-Weighted Probability Fusion

The recovery strategy $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q) \in [0,1]^q$ assigns a dimension-specific base preference to dependency-conditioned evidence, as defined in Section 3.4. For dimensions with a single observed class ($k_j = 1$), we fix $\theta_j = 1/2$ and restore an exact point-mass distribution after fusion.

For a stabilized probability vector $\boldsymbol{v} \in \mathbb{R}^{K_j}$, we compute its normalized entropy and reliability as

$$\mathsf{E}_j(\boldsymbol{v}) = \begin{cases} 0, & K_j = 1, \\ \mathrm{clip}_{[0,1]}\left(-\dfrac{\sum_{\ell=1}^{K_j} v_\ell \log v_\ell}{\log K_j} \right), & K_j > 1, \end{cases} \qquad r_j(\boldsymbol{v}) = \max\{1 - \mathsf{E}_j(\boldsymbol{v}), \delta\}, \tag{17}$$

where $\delta = 10^{-12}$ prevents numerical degeneracy. The clipping operation only removes floating-point excursions outside $[0, 1]$.

For an entry (i, j) with available dependency-conditioned evidence, let $r_{ij}^{\mathrm{F}} = r_j(\boldsymbol{u}_{ij}), r_{ij}^{\mathrm{D}} = r_j(\boldsymbol{c}_{ij})$, and define

$$d_{ij} = \theta_j r_{ij}^{\mathrm{D}} + (1 - \theta_j) r_{ij}^{\mathrm{F}}, \qquad w_{ij}(\boldsymbol{\theta}) = \begin{cases} \dfrac{\theta_j r_{ij}^{\mathrm{D}}}{d_{ij}}, & d_{ij} > \delta, \\ \theta_j, & d_{ij} \le \delta. \end{cases} \tag{18}$$

Thus, θ_j specifies the dimension-level preference, whereas r_{ij}^{F} and r_{ij}^{D} adapt this preference to the confidence of the two representations for each entry.

The stabilized logarithmic pooling used in the implementation is

$$a_{ij\ell} = (1 - w_{ij}) \log(u_{ij\ell} + \delta) + w_{ij} \log(c_{ij\ell} + \delta), \qquad \boldsymbol{p}_{ij}^{\boldsymbol{\theta}} = \mathrm{softmax}(\boldsymbol{a}_{ij}), \tag{19}$$

where $\boldsymbol{a}_{ij} = (a_{ij\ell})_{\ell=1}^{K_j}$. The softmax is evaluated using standard maximum subtraction for numerical stability.

Two boundary cases are handled explicitly. If dependency-conditioned evidence is unavailable, CSER sets $w_{ij} = 0$ and directly returns $\boldsymbol{p}_{ij}^{\boldsymbol{\theta}} = \boldsymbol{u}_{ij}$. If $k_j = 1$, the final distribution is replaced by $\boldsymbol{e}_{\ell_j}$, the point mass on the sole observed class ℓ_j. Consequently, every genuinely missing entry receives a valid probability distribution. CSER does not reject recovery targets according to confidence.

A.3 Constrained Pseudo-Missing Pattern Generation

This subsection specifies the feasible integer quotas underlying the pseudo-missing constraint in Section 3.5. Based on the observed class groups $\mathcal{I}_{j\ell}^{o}$, their sizes $n_{j\ell}$, and the empirical missing rate ρ_j. To preserve at least one observed instance of every represented class, the masking capacity of class ℓ is

$$b_{j\ell} = (n_{j\ell} - 1)_+, \qquad (a)_+ = \max(a, 0). \tag{20}$$

The resulting number of pseudo-missing entries in dimension j is

$$h_j = \begin{cases} 0, & k_j = 1, \\ \min\left\{ \left\lfloor \rho_j |\mathcal{I}_j^o| + \frac{1}{2} \right\rfloor, \sum_{\ell=1}^{K_j} b_{j\ell} \right\}, & k_j > 1, \end{cases} \tag{21}$$

where k_j is the number of classes represented among the observed labels of dimension j. Thus, the nominal missing rate determines the target challenge size, while the class-preservation constraint reduces it whenever the observed support is insufficient. A dimension with only one observed class is not used to generate pseudo-missing challenges.

The dimension-wise quota h_j is distributed across classes according to their observed support. Let

$$\eta_{j\ell} = \rho_j n_{j\ell}, \qquad h_{j\ell}^{(0)} = \min\{\lfloor \eta_{j\ell} \rfloor, b_{j\ell}\}. \tag{22}$$

Starting from $h_{j\ell}^{(0)}$, residual entries are allocated according to the fractional parts of $\eta_{j\ell}$, subject to the capacities $b_{j\ell}$, until the dimension-wise target h_j is reached. Fractional-part ties are resolved by class index. The final quotas therefore satisfy

$$\sum_{\ell=1}^{K_j} h_{j\ell} = h_j, \qquad 0 \le h_{j\ell} \le b_{j\ell}. \tag{23}$$

The candidate groups and class-wise quotas are fixed before co-evolution. Each challenge is represented by $H_{j\ell} \subseteq \mathcal{I}_{j\ell}^{o}$ with $|H_{j\ell}| = h_{j\ell}$, and their union defines the pseudo-missing set $H \subseteq \Omega^o$. The corresponding challenge mask is used only to identify recovery targets for fitness evaluation and is distinct from the original observation mask $\mathbf{M}$. Algorithm A2 summarizes the initialization and quota-preserving variation of feasible pseudo-missing challenges.

Algorithm A2 Initialization and Quota-Preserving Variation of Pseudo-Missing Challenges

Require: Observed class groups $\{\mathcal{I}^o_{j\ell}\}$, class-wise quotas $\{h_{j\ell}\}$, crossover probability p_c, mutation probability p_m
Ensure: Feasible pseudo-missing challenges satisfying all quotas
Initialization:
for each active dimension j and represented class ℓ **do**
  Sample $h_{j\ell}$ indices uniformly without replacement from $\mathcal{I}^o_{j\ell}$ to form $H_{j\ell}$.
end for
Form $H = \bigcup_{j,\ell}\{(i,j) : i \in H_{j\ell}\}$.
Variation:
Copy two selected parent challenges $H^{(1)}$ and $H^{(2)}$ to their offspring.
if a crossover draw is below p_c **then**
  for each offspring and each group (j,ℓ) **do**
    Retain the parental intersection $I = H^{(1)}_{j\ell} \cap H^{(2)}_{j\ell}$.
    Sample additional entries from $(H^{(1)}_{j\ell} \cup H^{(2)}_{j\ell}) \setminus I$ without replacement until possible.
    If necessary, complete the remaining quota from $\mathcal{I}^o_{j\ell}$ outside the current offspring set.
  end for
end if
for each offspring **do**
  if a mutation draw is below p_m **then**
    Select uniformly a group (j,ℓ) satisfying $0 < h_{j\ell} < |\mathcal{I}^o_{j\ell}|$.
    Swap one selected and one unselected index within that group.
  end if
  Form the corresponding challenge mask and verify $|H_{j\ell}| = h_{j\ell}$ for all (j,ℓ).
end for
return Feasible challenge offspring.

A.4 Fitness Evaluation, Sequential Co-Evolution, and Elite Archives

Let $\mathcal{R} = \{\boldsymbol{\theta}^{(r)}\}_{r=1}^{N_R}$ and $\mathcal{H} = \{H^{(a)}\}_{a=1}^{N_H}$ denote the current recovery-strategy and challenge populations. Their interactions are summarized by the pairwise recovery-loss matrix

$$\mathbf{Q}(\mathcal{R},\mathcal{H}) = [Q_{ra}], \qquad Q_{ra} = L(\boldsymbol{\theta}^{(r)}, H^{(a)}). \tag{24}$$

For each recovery strategy, the fixed probability representations are fused once to obtain the entry-wise losses $e_{ij}(\boldsymbol{\theta})$ on observed labels. A challenge then selects a sparse subset of these losses, assigning weight $1/(|\mathcal{J}|h_j)$ to each selected entry in active dimension j. Hence, the row and column means of $\mathbf{Q}$ directly yield the recovery and challenge fitness values in Eq. 10, respectively.

The two populations are updated sequentially. Recovery strategies are first optimized against the current challenge population by minimizing mean recovery loss. The resulting recovery population is then used to update the challenge population, whose objective is to maximize the same loss and thereby expose remaining recovery weaknesses. After both updates, the two populations are re-evaluated jointly, and their best individuals under the respective mean-fitness objectives are added to the elite archives. Algorithm A3 summarizes the sequential updates of the recovery and challenge populations together with generation-wise elite archiving.

For recovery-strategy variation, arithmetic crossover is applied to each selected parent pair with probability p_c. For coordinate j, independent $\lambda_j \sim \mathrm{Uniform}[0,1]$ produces complementary offspring

$$\lambda_j\theta^{(1)}_j + (1-\lambda_j)\theta^{(2)}_j \quad \text{and} \quad \lambda_j\theta^{(2)}_j + (1-\lambda_j)\theta^{(1)}_j.$$

Each offspring coordinate is subsequently perturbed by $\mathcal{N}(0,\sigma_m^2)$ with probability p_m, clipped to $[0,1]$, and reset to $1/2$ when $k_j = 1$.

Both populations use tournament selection and parent–offspring environmental selection under their respective optimization directions. Fitness ties preserve the earlier candidate in the corresponding ordered selection list. The elite archives contain only the generation-wise best individuals after both

Algorithm A3 Sequential Competitive Co-Evolution and Elite Archiving

Require: Fixed probability representations and caches, $\mathbf{Y}_{\Omega^o}$, feasible challenge sets and quotas, population sizes N_R, N_H, generations T
Ensure: Initial and final populations and elite archives $\mathcal{A}_R, \mathcal{A}_H$
if $\mathcal{J} = \varnothing$ **then**
    Set $\boldsymbol{\theta} = (1/2, \ldots, 1/2)$ and $H = \varnothing$; retain them as singleton initial and final populations; **return**
end if
Initialize $\mathcal{R}^0$ by sampling each free coordinate independently from $[0, 1]$; fix coordinates with $k_j = 1$ to $1/2$.
Initialize $\mathcal{H}^0$ using the constrained challenge construction in Algorithm A2.
Initialize $\mathcal{A}_R = \varnothing$ and $\mathcal{A}_H = \varnothing$.
for $t = 1, \ldots, T$ **do**
    Evaluate $\mathbf{Q}(\mathcal{R}^{t-1}, \mathcal{H}^{t-1})$ and compute recovery fitness values.
    Generate recovery offspring by tournament selection, arithmetic crossover, and Gaussian mutation.
    Evaluate the offspring against $\mathcal{H}^{t-1}$ and retain the N_R strategies with the lowest mean recovery loss as $\mathcal{R}^t$.
    Recompute challenge fitness against the updated recovery population $\mathcal{R}^t$.
    Generate challenge offspring by tournament selection and the quota-preserving variation operators in Algorithm A2.
    Evaluate the offspring against $\mathcal{R}^t$ and retain the N_H challenges with the highest mean recovery loss as $\mathcal{H}^t$.
    Recompute $\mathbf{Q}(\mathcal{R}^t, \mathcal{H}^t)$ and both populations' mean fitness values.
    Append the best recovery strategy and challenge to $\mathcal{A}_R$ and $\mathcal{A}_H$, respectively.
end for
return $\mathcal{R}^0, \mathcal{H}^0, \mathcal{R}^T, \mathcal{H}^T, \mathcal{A}_R, \mathcal{A}_H$.

populations have been updated and re-evaluated. Evolution runs for a fixed number of generations, while the probability representations remain unchanged throughout the search.

A.5 Archive-Based Robust Selection and Multiple Imputation

After evolution, CSER forms two historical candidate pools from the initial populations, generation-wise elite archives, and final populations:

$$\mathcal{R}_{\text{all}} = \mathcal{R}^0 \cup \mathcal{A}_R \cup \mathcal{R}^T, \qquad \mathcal{H}_{\text{all}} = \mathcal{H}^0 \cup \mathcal{A}_H \cup \mathcal{H}^T.$$

Candidates are processed in this order and duplicate entries are removed while retaining their first occurrence. Two recovery strategies are treated as identical when their maximum coordinate difference is below 10^{-12}, whereas two challenges are identical when their binary masks coincide.

CSER then performs complete cross-play between the retained recovery strategies and challenges using the fixed probability representations. Consistent with Eq. 11, the final strategy $\boldsymbol{\theta}^\star$ is selected lexicographically by its maximum recovery loss, mean recovery loss, and candidate order. Thus, evolution uses population-average fitness to explore effective strategies, whereas the final historical selection favors a candidate with robust performance across the collected challenges. Algorithm A4 summarizes the archive-based selection of the final recovery strategy and its subsequent multiple-imputation procedure for downstream MDC learning.

Multiple imputation propagates the uncertainty of the recovered class distributions into downstream MDC learning rather than committing to a single hard completion. Observed labels remain unchanged, while missing labels are independently sampled across entries and imputations according to Eq. 12. For each imputation and dimension, the downstream classifier is trained independently on all completed training instances using only the original features. Predicted class probabilities are then averaged across imputations before the final class decision. The imputed labels are used only for downstream learning and never update the fixed probability representations or the co-evolutionary search.

Algorithm A4 Archive-Based Robust Selection and Multiple Imputation

Require: Fixed probability representations and caches, initial/final populations, elite archives, training data, query features, imputation count S
Ensure: Averaged class probabilities and MDC predictions
1: Assemble and deduplicate $\mathcal{R}_{\text{all}}$ and $\mathcal{H}_{\text{all}}$.
2: Evaluate all retained recovery–challenge pairs using the recovery loss in Eq. 9.
3: Select $\boldsymbol{\theta}^{\star}$ by minimum maximum loss, breaking ties by mean loss and then candidate order.
4: Obtain $\boldsymbol{p}_{ij}^{\boldsymbol{\theta}^{\star}}$ on Ω^m using Eq. 19.
5: **for** $s = 1, \ldots, S$ **do**
6: **for** $j = 1, \ldots, q$ **do**
7: Preserve observed labels and sample each missing label according to Eq. 12.
8: Fit the downstream classifier $g_j^{(s)}$ on all completed training instances using features alone.
9: Predict class probabilities for the query instances.
10: **end for**
11: **end for**
12: Average the predicted probabilities over the S imputations independently for each label dimension.
13: Predict the class with the largest averaged probability in each dimension.
14: **return** Averaged class probabilities and MDC predictions.

A.6 Implementation Details

The probability stages use $B = 3$ repetitions and a nominal 5 folds, subject to the support rules in Appendix A.1. We use $N_R = N_H = 50$, $T = 50$, tournament size 2, $p_c = 0.9$, $p_m = 0.1$, and $\sigma_m = 0.1$. Mutation probability is applied per coordinate for label recovery strategies and per individual for challenges. The archived benchmark runs use $S = 10$ imputations. Both estimation stages and the downstream models use a linear-kernel LIBSVM classifier with probability estimation enabled. The outer folds described above are separate from LIBSVM's internal probability estimation. Empty and single-class training sets use the constant predictors specified in Appendix A.1.

A.7 Computational Complexity

Let $K_\Sigma = \sum_{j=1}^{q} K_j,\ h = \sum_{j \in \mathcal{J}} h_j$, where K_Σ is the total number of class-probability components across all dimensions and h is the number of pseudo-missing entries contained in each challenge.

Probability representation construction. With B repetitions and at most F folds, the two-stage construction requires at most $2BqF$ fold-specific model fits. The feature-only stage uses d input features, whereas the dependency-conditioned model for dimension j uses $d + K_\Sigma - K_j + q - 1$ inputs. Since the computational cost of fitting these models depends on the underlying probabilistic classifier, we report the number of model fits rather than imposing a classifier-specific training complexity. Importantly, these estimators are constructed only once and are not retrained during co-evolution.

Competitive co-evolution. For one evaluation of N_R recovery strategies against N_H challenges, computing the fused probabilities and entry-wise recovery losses costs $O(N_R n K_\Sigma)$, while accumulating the losses over all recovery–challenge pairs costs $O(N_R N_H h)$. Therefore, ignoring lower-order selection and variation operations, T generations require

$$O(T\,[N_R n K_\Sigma + N_R N_H h])\,. \tag{25}$$

Archive-based selection. The historical candidate pools contain at most $\widehat{N}_R = 2N_R + T$ recovery strategies and $\widehat{N}_H = 2N_H + T$ challenges before deduplication. Therefore, the final complete cross-play requires

$$O\Big(\widehat{N}_R n K_\Sigma + \widehat{N}_R \widehat{N}_H h\Big) \tag{26}$$

operations for probability fusion and recovery-loss evaluation.

Multiple-imputation prediction on the MDC test set. After selecting $\boldsymbol{\theta}^\star$, CSER performs S independent imputations and trains one downstream classifier for each label dimension, resulting in Sq additional model fits. The fixed class-probability representations require $O(nK_\Sigma)$ memory, while the pairwise interaction matrix requires $O(N_R N_H)$ memory during evolution and $O(\widehat{N}_R \widehat{N}_H)$ memory during final cross-play.

B Implementation of Label-Imputation Strategies for Baseline Adaptation

As described in Section 4 in the main text, the fully supervised MDC baselines are adapted to incomplete-label training data using two preprocessing strategies. Both strategies operate only on the training data and preserve all entries in Ω^o. After imputation, the resulting completed label matrix is supplied to the corresponding MDC method, whose original classifier, hyperparameters, and training procedure remain unchanged.

Prototype-Based Label Imputation. Prototype-based imputation exploits feature-space proximity to infer missing labels. Given the training features $\mathbf{X} \in \mathbb{R}^{n\times d}$, the observed labels $\mathbf{Y}_{\Omega^o}$, and the observation mask $\mathbf{M}$, missing feature values, if present, are first replaced by their dimension-wise means and the resulting features are standardized to zero mean and unit variance, yielding $\widetilde{\mathbf{X}}$.

For dimension j, let

$$\mathcal{I}_j^m = \{i : (i,j) \in \Omega^m\}, \qquad \mathcal{C}_j^o = \{c \in \mathcal{C}_j : \exists\, i \in \mathcal{I}_j^o,\ y_{ij} = c\}.$$

For each observed class $c \in \mathcal{C}_j^o$, its feature prototype is defined as

$$\boldsymbol{\mu}_{jc} = \frac{1}{|\mathcal{I}_{jc}^o|} \sum_{i \in \mathcal{I}_{jc}^o} \widetilde{\boldsymbol{x}}_i, \qquad \mathcal{I}_{jc}^o = \{i \in \mathcal{I}_j^o : y_{ij} = c\}.$$

Each missing entry is assigned the class with the nearest prototype in squared Euclidean distance. Dimensions without missing entries require no update; if only one class is observed, all missing entries in that dimension are assigned that class. A dimension with no observed labels cannot be imputed by this strategy. Algorithm A5 summarizes the procedure.

Correlation-Based Label Imputation. Correlation-based imputation relies exclusively on dependencies among label dimensions and does not use the input features during label completion. Missing entries are first initialized by the observed mode of their corresponding dimension. From the observed labels, the method estimates a Laplace-smoothed marginal prior for each dimension and pairwise conditional probabilities between distinct dimensions.

For dimension j, let n_{ja} denote the number of observed entries with class a. With smoothing parameter $\alpha = 1$, the marginal log-prior is

$$\log P_j(a) = \log \frac{n_{ja} + \alpha}{\sum_{c \in \mathcal{C}_j} (n_{jc} + \alpha)}.$$

For an ordered pair (k, j), define the jointly observed set

$$\mathcal{S}_{kj} = \{i : (i,k) \in \Omega^o,\ (i,j) \in \Omega^o\}.$$

Pairs with $|\mathcal{S}_{kj}| < s_{\min}$ are discarded, where $s_{\min} = 5$. For the remaining pairs, let $n_{kj}(b,a)$ be the number of jointly observed samples with $y_{ik} = b$ and $y_{ij} = a$, and $n_{kj}(b) = \sum_a n_{kj}(b,a)$. The conditional probability is estimated as

$$\log P_{j|k}(a \mid b) = \log \frac{n_{kj}(b,a) + \alpha}{n_{kj}(b) + \alpha|\mathcal{C}_j|}.$$

The initialized labels are then refined iteratively. For each missing entry, the method selects the class maximizing its marginal log-prior plus the available conditional evidence from the other label dimensions. Observed labels remain fixed throughout the refinement. The procedure runs for at

most $T_{\mathrm{corr}} = 3$ sweeps and stops early when no missing label changes. Algorithm A6 summarizes the procedure.

For each baseline configuration, the completed matrix $\widetilde{\mathbf{Y}}$ is subsequently passed to DCOM, PIST, LEAD, or ECCO. The downstream MDC method retains its original feature processing, classifier, hyperparameter settings, and training procedure; therefore, the prototype- and correlation-based variants differ only in the label-imputation stage.

Algorithm A5 Prototype-Based Label Imputation

Require: Training data $(\mathbf{X}, \mathbf{Y}_{\Omega^o}, \mathbf{M})$
Ensure: Completed label matrix $\widetilde{\mathbf{Y}}$

Mean-impute missing feature values, if any, and standardize $\mathbf{X}$ to obtain $\widetilde{\mathbf{X}}$.
Initialize $\widetilde{\mathbf{Y}}$ with the observed labels.
for $j = 1, \ldots, q$ **do**
    Set $\mathcal{I}_j^m = \{i : (i, j) \in \Omega^m\}$.
    if $\mathcal{I}_j^m = \varnothing$ **then**
        continue
    end if
    Determine the observed class set $\mathcal{C}_j^o$ from $\mathcal{I}_j^o$.
    if $|\mathcal{C}_j^o| = 1$ **then**
        Assign the sole observed class to all $i \in \mathcal{I}_j^m$; **continue**.
    end if
    for each $c \in \mathcal{C}_j^o$ **do**
        Compute $\boldsymbol{\mu}_{jc} = |\mathcal{I}_{jc}^o|^{-1} \sum_{i \in \mathcal{I}_{jc}^o} \widetilde{\boldsymbol{x}}_i$.
    end for
    for each $i \in \mathcal{I}_j^m$ **do**
        $\widetilde{y}_{ij} \leftarrow \arg\min_{c \in \mathcal{C}_j^o} \|\widetilde{\boldsymbol{x}}_i - \boldsymbol{\mu}_{jc}\|_2^2$.
    end for
end for
Preserve $\widetilde{y}_{ij} = y_{ij}$ for all $(i, j) \in \Omega^o$.
return $\widetilde{\mathbf{Y}}$.

Algorithm A6 Correlation-Based Label Imputation

Require: Observed labels $\mathbf{Y}_{\Omega^o}$, observation mask $\mathbf{M}$, smoothing α, maximum sweeps T_{corr}, minimum pair support $s_{\min}$

Ensure: Completed label matrix $\widetilde{\mathbf{Y}}$

Initialize $\widetilde{\mathbf{Y}}$ with the observed labels.
for $j = 1, \ldots, q$ **do**
  Set $\mathcal{I}_j^m = \{i : (i,j) \in \Omega^m\}$.
  Compute the observed mode c_j^{mode} and initialize $\widetilde{y}_{ij} \leftarrow c_j^{\mathrm{mode}}$ for all $i \in \mathcal{I}_j^m$.
  Estimate $\log P_j(a)$ for all $a \in \mathcal{C}_j$ using Laplace smoothing.
end for
for each ordered pair (k, j) with $k \neq j$ **do**
  Construct $\mathcal{S}_{kj} = \{i : (i,k) \in \Omega^o,\ (i,j) \in \Omega^o\}$.
  if $|\mathcal{S}_{kj}| < s_{\min}$ **then**
    continue
  end if
  Estimate $\log P_{j|k}(a \mid b)$ for $b \in \mathcal{C}_k$ and $a \in \mathcal{C}_j$.
end for
for $t = 1, \ldots, T_{\mathrm{corr}}$ **do**
  Set changed $\leftarrow$ false.
  for $j = 1, \ldots, q$ **do**
    for each $i \in \mathcal{I}_j^m$ **do**
      Update

$$\widetilde{y}_{ij} \leftarrow \arg\max_{a \in \mathcal{C}_j} \left[\log P_j(a) + \sum_{\substack{k \neq j: \\ |\mathcal{S}_{kj}| \geq s_{\min}}} \log P_{j|k}\big(a \mid \widetilde{y}_{ik}\big) \right].$$

      Record whether the label changed.
    end for
  end for
  Restore $\widetilde{y}_{ij} = y_{ij}$ for all $(i,j) \in \Omega^o$.
  if no missing label changed **then**
    break
  end if
end for
return $\widetilde{\mathbf{Y}}$.

C FULL COMPARISON OF TEST RESULTS ON ALL EVALUATION METRICS AND RUNTIME

C.1 REPORTING PROTOCOL AND EVALUATION RESULTS.

This section provides the complete test results complementary to the main-text analysis. Statistical significance in Table 2 is assessed using paired Wilcoxon signed-rank tests on dataset-level mean scores, pairing CSER and each baseline on the same datasets. For each missing rate and evaluation metric, Holm's step-down correction is applied across the eight baseline comparisons to control the family-wise Type I error rate at $\alpha = 0.05$. This adjustment accounts for the increased risk of false-positive findings when multiple pairwise tests are performed. Tables S-I and S-II report Hamming Score (HS) under 30% and 50% training-label missingness, respectively, while the corresponding 70% result is presented in Table 3 of the main text. Tables S-III, S-V, and S-VII report Exact Match (EM) under 30%, 50%, and 70% missingness, respectively, and Tables S-IV, S-VI, and S-VIII report Sub-Exact Match (SEM) under the same settings. Each predictive result is reported as the mean $\pm$ standard deviation over five repeated runs with different missing-label realizations at each missing rate. The best mean value is highlighted in bold and the second-best is underlined, and the last row reports the average rank across datasets. Tables S-IX–S-XI further report the mean running time under the three missing-label rates.

Detailed performance analysis. The complete results reveal a consistent trend across the three evaluation metrics. In particular, the relative advantage of CSER becomes stronger as the training supervision becomes more incomplete. Its average HS rank improves from 1.588 at 30% missingness to 1.265 at 50%, and reaches 1.000 at 70%. A similar pattern is observed for the stricter joint-prediction metrics: the average EM rank improves from 2.029 to 1.618 and 1.323, while the corresponding SEM ranks are 2.265, 1.618, and 1.588. These results show that the benefit of CSER is not restricted to dimension-wise prediction accuracy, but is also maintained when multiple class dimensions must be predicted correctly simultaneously.

The substantial performance gaps under severe missingness can be interpreted in light of how the compared baselines handle uncertain training labels. The adapted MDC baselines produce a single completed label matrix before downstream learning. At 70% missingness, imputed label entries constitute the majority of training-label entries, so inaccurate completion can substantially affect the supervision available to the downstream classifier. Errors in the completed labels may also distort the cross-dimensional relationships used by dependency-aware learners. Consequently, a strong MDC learner trained on complete labels may still perform poorly when its training supervisions are dominated by uncertain imputations. This provides a plausible explanation for the large gaps, particularly when limited observed supervision makes completion difficult.

CSER addresses this difficulty by retaining feature-only and dependency-conditioned class-probability representations and learning their relative contributions through dimension-wise preferences and instance-specific confidence. Dependency-conditioned evidence receives less weight when its entropy-based confidence is low relative to the feature-only evidence. Recovery strategies are evaluated on dynamically selected observed-label targets and finally compared against collected historical challenges. Multiple imputation further allows downstream prediction to average over alternative label completions instead of depending on a single completed matrix. Together, these mechanisms provide a plausible explanation for the favorable performance under severe missingness. The observed gaps nevertheless reflect comparisons between complete learning pipelines, including their recovery procedures and downstream learners, rather than isolating the effect of any single component.

The advantage is especially visible on several challenging datasets under 70% missingness. For example, CSER achieves HS/EM/SEM values of $0.897/0.800/0.994$ on Voice, $0.898/0.589/0.908$ on CoIL2000, and $0.897/0.712/0.980$ on TIC2000. The gains on EM are particularly informative because this metric requires all class dimensions of an unseen instance to be predicted correctly, making it sensitive to errors propagated across dimensions. Together with the corresponding HS and SEM results, these outcomes indicate that the recovered probabilistic supervision remains useful not only for individual dimensions but also for joint multi-dimensional prediction. Since all three metrics are computed exclusively on unseen test instances, the results provide a direct evidence that the supervision recovered from incomplete training labels transfers to inductive MDC prediction. At

30% missingness, several dataset-level results are matched or occasionally exceeded by individual baseline variants, indicating that the advantage of CSER is not based on uniform dominance in every low-missingness case. Nevertheless, its favorable average ranks and the Holm-adjusted tests in the main text show that the overall advantage is maintained across datasets, while becoming substantially clearer as the missing-label rate increases.

Table S-I: Test Hamming Score (HS) under 30% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.949±0.000**	0.942±0.003	0.947±0.003	**0.949±0.000**	**0.949±0.000**	0.886±0.016	0.922±0.007	0.941±0.010	0.948±0.004
Jura	**0.666±0.014**	0.578±0.039	0.574±0.030	0.577±0.023	0.557±0.029	0.607±0.031	0.609±0.009	0.580±0.071	0.511±0.019
Enb	**0.798±0.023**	0.701±0.013	0.697±0.023	0.677±0.008	0.652±0.012	0.745±0.013	0.719±0.028	0.692±0.025	0.652±0.033
Song	**0.768±0.008**	0.697±0.004	0.749±0.014	0.743±0.018	0.766±0.009	0.767±0.013	**0.768±0.007**	0.765±0.011	0.762±0.012
WQplants	**0.705±0.003**	0.675±0.015	0.703±0.000	0.703±0.001	0.703±0.001	0.643±0.007	0.696±0.003	0.700±0.003	0.693±0.023
WQanimals	0.658±0.002	0.659±0.001	0.660±0.000	0.660±0.000	0.660±0.000	0.555±0.008	0.652±0.005	0.616±0.027	**0.661±0.001**
WaterQuality	**0.691±0.001**	0.681±0.013	0.690±0.000	0.688±0.004	0.690±0.000	0.613±0.012	0.687±0.002	0.667±0.015	0.690±0.000
BeLaE	**0.433±0.003**	0.426±0.012	0.409±0.008	0.376±0.003	0.360±0.004	0.410±0.011	0.369±0.006	0.414±0.026	0.358±0.007
Voice	**0.916±0.010**	0.801±0.015	0.829±0.014	0.668±0.010	0.622±0.014	0.750±0.014	0.729±0.027	0.680±0.059	0.574±0.145
Rf1	0.748±0.003	0.733±0.024	**0.788±0.008**	0.720±0.011	0.743±0.004	0.663±0.017	0.676±0.018	0.685±0.025	0.689±0.077
Thyroid	**0.971±0.000**	0.957±0.001	0.962±0.000	0.958±0.000	0.957±0.000	0.956±0.001	0.960±0.000	0.962±0.003	0.960±0.003
Pain	**0.960±0.000**	0.958±0.000	0.959±0.000	0.958±0.000	0.958±0.000	0.956±0.001	**0.960±0.000**	0.959±0.001	0.958±0.000
CoIL2000	**0.932±0.002**	0.906±0.002	0.898±0.006	0.856±0.002	0.759±0.012	0.848±0.002	0.818±0.001	0.685±0.051	0.622±0.036
TIC2000	**0.926±0.002**	0.904±0.002	0.912±0.005	0.902±0.002	0.850±0.021	0.830±0.005	0.828±0.004	0.754±0.055	0.729±0.048
Fera	0.682±0.002	0.609±0.001	0.601±0.004	0.599±0.004	0.580±0.003	**0.713±0.002**	0.709±0.003	0.584±0.013	0.554±0.008
Adult	**0.710±0.001**	0.700±0.002	0.697±0.001	0.693±0.001	0.649±0.002	0.681±0.001	0.662±0.002	0.563±0.050	0.517±0.009
Default	**0.666±0.000**	0.637±0.000	0.639±0.000	0.633±0.000	0.640±0.000	0.428±0.000	0.464±0.000	0.603±0.001	0.603±0.001
Avg. Rank	**1.588**	4.824	3.706	5.147	5.500	6.177	5.206	6.029	6.824

Table S-II: Test Hamming Score (HS) under 50% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.949±0.000**	0.939±0.002	0.936±0.007	0.945±0.006	**0.949±0.000**	0.868±0.029	0.937±0.006	0.904±0.031	0.947±0.002
Jura	**0.672±0.029**	0.566±0.062	0.500±0.007	0.582±0.027	0.578±0.013	0.531±0.042	0.530±0.006	0.543±0.032	0.495±0.003
Enb	**0.799±0.012**	0.679±0.011	0.598±0.075	0.674±0.014	0.631±0.032	0.682±0.019	0.618±0.061	0.657±0.038	0.536±0.115
Song	**0.767±0.006**	0.674±0.022	0.724±0.009	0.747±0.011	0.746±0.008	0.708±0.012	0.728±0.009	0.706±0.030	0.720±0.011
WQplants	**0.706±0.001**	0.692±0.017	0.704±0.000	0.703±0.000	0.704±0.000	0.558±0.025	0.703±0.001	0.632±0.029	0.704±0.000
WQanimals	0.656±0.004	0.639±0.015	**0.660±0.000**	0.618±0.009	**0.660±0.000**	0.462±0.008	0.658±0.002	0.525±0.042	**0.660±0.000**
WaterQuality	**0.691±0.001**	0.670±0.012	0.690±0.000	0.679±0.013	0.690±0.000	0.519±0.012	0.689±0.001	0.546±0.016	0.690±0.000
BeLaE	**0.418±0.008**	0.380±0.010	0.371±0.007	0.349±0.003	0.354±0.005	0.384±0.010	0.359±0.006	0.369±0.041	0.356±0.005
Voice	**0.911±0.004**	0.724±0.026	0.651±0.028	0.612±0.016	0.521±0.007	0.706±0.020	0.585±0.033	0.598±0.158	0.461±0.113
Rf1	**0.745±0.010**	0.655±0.021	0.723±0.010	0.637±0.015	0.685±0.010	0.614±0.025	0.645±0.022	0.614±0.066	0.550±0.114
Thyroid	**0.971±0.001**	0.957±0.000	0.959±0.001	0.957±0.000	0.957±0.000	0.926±0.002	0.959±0.001	0.953±0.005	0.958±0.000
Pain	**0.960±0.000**	0.958±0.000	0.958±0.000	0.958±0.000	0.958±0.000	0.911±0.004	0.959±0.000	0.941±0.006	0.958±0.000
CoIL2000	**0.918±0.002**	0.846±0.007	0.773±0.011	0.806±0.006	0.630±0.004	0.795±0.004	0.721±0.001	0.658±0.037	0.595±0.013
TIC2000	**0.914±0.003**	0.849±0.003	0.843±0.008	0.846±0.005	0.774±0.014	0.781±0.003	0.770±0.002	0.757±0.049	0.687±0.001
Fera	**0.675±0.001**	0.556±0.003	0.537±0.003	0.569±0.003	0.532±0.000	0.572±0.004	0.599±0.004	0.510±0.022	0.532±0.000
Adult	**0.707±0.002**	0.667±0.003	0.598±0.007	0.672±0.002	0.572±0.010	0.647±0.004	0.584±0.004	0.543±0.023	0.518±0.005
Default	**0.666±0.001**	0.574±0.000	0.609±0.000	0.579±0.000	0.604±0.000	0.413±0.000	0.464±0.000	0.603±0.001	0.603±0.001
Avg. Rank	**1.265**	4.706	4.441	4.677	4.882	6.265	5.294	7.118	6.353

Table S-III: Test Exact Match (EM) under 30% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.877±0.000**	0.855±0.008	0.871±0.008	**0.877±0.000**	**0.877±0.000**	0.778±0.014	0.818±0.023	0.865±0.015	**0.877±0.000**
Jura	**0.451±0.021**	0.346±0.045	0.289±0.042	0.297±0.014	0.273±0.020	0.381±0.037	0.349±0.018	0.324±0.101	0.195±0.041
Enb	**0.596±0.045**	0.403±0.026	0.399±0.043	0.353±0.016	0.306±0.019	0.491±0.026	0.444±0.045	0.385±0.047	0.322±0.039
Song	**0.466±0.016**	0.342±0.020	0.424±0.018	0.410±0.045	0.459±0.011	0.463±0.023	0.453±0.006	0.459±0.023	0.451±0.011
WQplants	**0.171±0.003**	0.101±0.044	0.168±0.002	0.169±0.000	0.169±0.000	0.099±0.035	0.164±0.003	0.169±0.006	0.143±0.059
WQanimals	**0.094±0.002**	0.088±0.002	0.089±0.000	0.089±0.000	0.089±0.000	0.010±0.004	0.084±0.009	0.031±0.031	0.091±0.004
WaterQuality	**0.028±0.000**	0.018±0.014	**0.028±0.000**	0.025±0.006	**0.028±0.000**	0.008±0.009	**0.028±0.000**	0.008±0.014	**0.028±0.000**
BeLaE	**0.029±0.005**	0.024±0.002	0.022±0.003	0.020±0.004	0.019±0.006	0.016±0.003	0.022±0.006	0.022±0.006	0.021±0.007
Voice	**0.833±0.019**	0.615±0.027	0.674±0.025	0.410±0.008	0.456±0.025	0.540±0.026	0.523±0.036	0.421±0.120	0.326±0.226
Rf1	0.115±0.015	0.101±0.021	**0.154±0.018**	0.125±0.022	0.122±0.003	0.084±0.019	0.040±0.016	0.065±0.042	0.060±0.058
Thyroid	**0.816±0.004**	0.731±0.010	0.760±0.003	0.727±0.002	0.725±0.000	0.732±0.007	0.747±0.002	0.768±0.024	0.746±0.018
Pain	0.792±0.002	0.797±0.000	0.795±0.001	0.797±0.000	0.797±0.000	0.756±0.008	**0.798±0.002**	0.795±0.001	0.797±0.000
CoIL2000	**0.718±0.008**	0.627±0.006	0.612±0.021	0.465±0.007	0.229±0.016	0.476±0.006	0.397±0.006	0.157±0.058	0.091±0.017
TIC2000	**0.791±0.003**	0.735±0.006	0.758±0.012	0.727±0.004	0.589±0.052	0.568±0.012	0.567±0.012	0.411±0.094	0.365±0.063
Fera	0.247±0.002	0.195±0.003	0.194±0.002	0.193±0.003	0.193±0.002	0.282±0.003	**0.310±0.008**	0.177±0.012	0.187±0.001
Adult	0.243±0.002	**0.262±0.001**	0.213±0.005	0.253±0.002	0.119±0.001	0.238±0.004	0.187±0.005	0.078±0.062	0.050±0.059
Default	**0.174±0.001**	0.140±0.000	0.142±0.000	0.133±0.000	0.144±0.000	0.032±0.000	0.046±0.000	0.115±0.002	0.115±0.002
Avg. Rank	**2.029**	4.676	4.147	5.147	5.294	5.912	5.059	6.235	6.500

Table S-IV: Test Sub-Exact Match (SEM) under 30% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.969±0.000**	**0.969±0.000**	**0.969±0.000**	**0.969±0.000**	**0.969±0.000**	0.902±0.028	0.948±0.008	0.958±0.015	**0.969±0.000**
Jura	**0.881±0.011**	0.811±0.053	0.859±0.031	0.857±0.035	0.841±0.041	0.832±0.042	0.870±0.015	0.835±0.052	0.827±0.020
Enb	**1.000±0.000**	**1.000±0.000**	0.996±0.006	**1.000±0.000**	0.997±0.006	**1.000±0.000**	0.995±0.012	0.998±0.003	0.982±0.041
Song	0.847±0.015	0.773±0.022	0.832±0.030	0.827±0.011	0.851±0.020	0.848±0.013	**0.863±0.018**	0.841±0.018	0.843±0.030
WQplants	0.395±0.011	0.334±0.040	0.404±0.000	0.403±0.002	**0.406±0.004**	0.278±0.023	0.376±0.008	0.384±0.014	0.386±0.040
WQanimals	0.268±0.011	0.274±0.003	0.276±0.000	0.276±0.000	0.276±0.000	0.102±0.013	0.261±0.002	0.179±0.072	**0.279±0.004**
WaterQuality	**0.116±0.004**	0.084±0.033	0.108±0.000	0.103±0.010	0.108±0.000	0.032±0.021	0.108±0.003	0.056±0.029	0.108±0.000
BeLaE	**0.140±0.005**	0.126±0.014	0.126±0.012	0.087±0.003	0.101±0.011	0.110±0.012	0.102±0.010	0.112±0.006	0.102±0.007
Voice	**0.999±0.001**	0.987±0.004	0.984±0.005	0.926±0.011	0.789±0.008	0.960±0.004	0.936±0.022	0.939±0.050	0.822±0.082
Rf1	0.333±0.011	0.340±0.036	**0.500±0.014**	0.338±0.026	0.392±0.012	0.281±0.010	0.223±0.027	0.272±0.043	0.264±0.111
Thyroid	**0.982±0.001**	0.972±0.006	0.976±0.001	0.977±0.001	0.977±0.000	0.965±0.003	0.976±0.000	0.976±0.002	0.976±0.000
Pain	**0.903±0.001**	0.894±0.000	0.895±0.000	0.894±0.000	0.894±0.000	**0.903±0.004**	0.902±0.003	0.895±0.003	0.894±0.000
CoIL2000	**0.945±0.004**	0.914±0.005	0.898±0.011	0.842±0.004	0.656±0.031	0.824±0.006	0.768±0.002	0.493±0.096	0.352±0.050
TIC2000	**0.987±0.002**	0.978±0.002	0.980±0.005	0.978±0.001	0.961±0.011	0.924±0.003	0.923±0.005	0.861±0.065	0.830±0.076
Fera	0.507±0.004	0.401±0.003	0.392±0.005	0.384±0.007	0.369±0.004	**0.565±0.006**	0.556±0.007	0.370±0.022	0.338±0.006
Adult	**0.669±0.003**	0.650±0.005	0.650±0.005	0.636±0.002	0.578±0.004	0.619±0.004	0.586±0.003	0.452±0.059	0.388±0.015
Default	**0.591±0.001**	0.540±0.008	0.544±0.005	0.530±0.000	0.546±0.000	0.210±0.010	0.261±0.004	0.479±0.008	0.479±0.008
Avg. Rank	**2.265**	4.882	3.588	4.853	4.824	5.882	5.618	6.441	6.647

Table S-V: Test Exact Match (EM) under 50% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.877±0.000**	0.849±0.007	0.840±0.021	0.871±0.008	**0.877±0.000**	0.735±0.063	0.849±0.020	0.794±0.090	0.874±0.007
Jura	**0.449±0.048**	0.316±0.072	0.176±0.014	0.289±0.041	0.292±0.007	0.295±0.078	0.235±0.007	0.305±0.037	0.165±0.006
Enb	**0.599±0.025**	0.357±0.022	0.281±0.060	0.348±0.027	0.284±0.048	0.365±0.037	0.317±0.042	0.329±0.060	0.225±0.059
Song	**0.470±0.019**	0.262±0.055	0.394±0.018	0.433±0.026	0.437±0.011	0.328±0.027	0.411±0.014	0.334±0.057	0.395±0.020
WQplants	**0.173±0.004**	0.120±0.065	0.169±0.000	0.165±0.008	0.169±0.000	0.008±0.004	0.172±0.003	0.036±0.034	0.169±0.000
WQanimals	**0.091±0.004**	0.052±0.026	0.089±0.000	0.023±0.006	0.089±0.000	0.004±0.004	0.089±0.000	0.016±0.018	0.089±0.000
WaterQuality	**0.028±0.000**	0.006±0.013	**0.028±0.000**	0.017±0.015	**0.028±0.000**	0.000±0.000	**0.028±0.000**	0.000±0.000	**0.028±0.000**
BeLaE	**0.026±0.004**	0.015±0.003	0.023±0.006	0.010±0.007	0.021±0.006	0.015±0.006	0.022±0.006	0.012±0.006	0.021±0.006
Voice	**0.826±0.007**	0.482±0.049	0.418±0.043	0.313±0.031	0.263±0.012	0.464±0.034	0.323±0.047	0.353±0.201	0.203±0.147
Rf1	0.092±0.006	0.023±0.015	**0.128±0.010**	0.019±0.021	0.119±0.000	0.040±0.036	0.028±0.022	0.020±0.023	0.024±0.053
Thyroid	**0.812±0.006**	0.725±0.000	0.740±0.005	0.725±0.001	0.725±0.000	0.566±0.010	0.735±0.004	0.710±0.020	0.726±0.003
Pain	0.794±0.001	**0.797±0.000**	0.796±0.000	**0.797±0.000**	**0.797±0.000**	0.552±0.031	**0.797±0.001**	0.697±0.024	**0.797±0.000**
CoIL2000	**0.662±0.009**	0.448±0.019	0.286±0.029	0.346±0.008	0.095±0.003	0.350±0.004	0.208±0.006	0.130±0.035	0.077±0.008
TIC2000	**0.759±0.010**	0.604±0.007	0.596±0.020	0.590±0.014	0.428±0.022	0.476±0.005	0.452±0.006	0.416±0.090	0.313±0.002
Fera	**0.241±0.000**	0.157±0.011	0.186±0.000	0.186±0.002	0.186±0.000	0.092±0.006	0.211±0.001	0.099±0.025	0.186±0.000
Adult	**0.239±0.004**	0.219±0.004	0.136±0.005	0.229±0.003	0.115±0.000	0.192±0.007	0.130±0.002	0.127±0.020	0.038±0.051
Default	**0.174±0.001**	0.091±0.000	0.119±0.000	0.091±0.000	0.116±0.000	0.027±0.000	0.046±0.000	0.115±0.002	0.115±0.002
Avg. Rank	**1.618**	4.971	4.353	5.353	4.706	6.529	4.647	7.000	5.824

Table S-VI: Test Sub-Exact Match (SEM) under 50% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.969±0.000**	**0.969±0.000**	**0.969±0.000**	0.963±0.008	**0.969±0.000**	0.895±0.030	0.963±0.008	0.929±0.014	**0.969±0.000**
Jura	**0.895±0.026**	0.816±0.055	0.824±0.000	0.876±0.015	0.865±0.019	0.768±0.020	0.824±0.014	0.781±0.036	0.824±0.000
Enb	**1.000±0.000**	**1.000±0.000**	0.916±0.101	**1.000±0.000**	0.977±0.025	**1.000±0.000**	0.919±0.090	0.986±0.032	0.848±0.171
Song	**0.839±0.011**	0.778±0.025	0.785±0.025	0.815±0.008	0.809±0.014	0.816±0.017	0.780±0.020	0.804±0.043	0.775±0.016
WQplants	**0.405±0.004**	0.369±0.043	0.404±0.000	0.404±0.000	0.404±0.000	0.099±0.042	0.401±0.004	0.236±0.084	0.404±0.000
WQanimals	0.259±0.018	0.241±0.028	**0.276±0.000**	0.206±0.023	**0.276±0.000**	0.039±0.014	0.264±0.008	0.087±0.040	**0.276±0.000**
WaterQuality	**0.108±0.006**	0.052±0.033	**0.108±0.000**	0.080±0.034	**0.108±0.000**	0.000±0.000	**0.108±0.000**	0.001±0.002	**0.108±0.000**
BeLaE	**0.134±0.010**	0.093±0.015	0.106±0.018	0.064±0.008	0.098±0.016	0.081±0.008	0.099±0.020	0.090±0.032	0.097±0.017
Voice	**0.997±0.001**	0.966±0.010	0.883±0.025	0.911±0.003	0.778±0.002	0.948±0.012	0.847±0.023	0.844±0.115	0.719±0.089
Rf1	0.345±0.027	0.225±0.068	**0.422±0.023**	0.186±0.040	0.337±0.007	0.197±0.037	0.154±0.037	0.169±0.091	0.122±0.150
Thyroid	**0.982±0.001**	0.977±0.000	0.976±0.000	0.976±0.000	0.977±0.000	0.922±0.007	0.976±0.001	0.963±0.016	0.977±0.000
Pain	**0.898±0.003**	0.894±0.000	0.894±0.000	0.894±0.000	0.894±0.000	0.780±0.019	0.895±0.000	0.850±0.026	0.894±0.000
CoIL2000	**0.932±0.006**	0.818±0.012	0.674±0.020	0.744±0.015	0.364±0.006	0.724±0.012	0.570±0.009	0.442±0.072	0.314±0.027
TIC2000	**0.984±0.002**	0.945±0.004	0.937±0.006	0.951±0.004	0.898±0.025	0.875±0.009	0.866±0.006	0.864±0.052	0.764±0.002
Fera	**0.498±0.005**	0.341±0.003	0.326±0.000	0.356±0.004	0.326±0.000	0.331±0.005	0.394±0.006	0.273±0.032	0.326±0.001
Adult	**0.662±0.005**	0.598±0.006	0.472±0.007	0.603±0.005	0.417±0.017	0.565±0.005	0.451±0.006	0.396±0.044	0.385±0.015
Default	**0.594±0.003**	0.419±0.018	0.489±0.008	0.431±0.000	0.481±0.000	0.187±0.009	0.259±0.010	0.479±0.008	0.479±0.008
Avg. Rank	**1.618**	4.676	4.294	4.500	4.588	6.559	5.441	7.265	6.059

Table S-VII: Test Exact Match (EM) under 70% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.877±0.000**	0.840±0.018	0.843±0.007	0.871±0.008	0.874±0.007	0.671±0.172	0.871±0.008	0.788±0.166	0.870±0.000
Jura	**0.389±0.031**	0.251±0.020	0.151±0.041	0.127±0.041	0.173±0.077	0.189±0.060	0.173±0.061	0.205±0.065	0.135±0.051
Enb	**0.521±0.088**	0.361±0.033	0.230±0.080	0.370±0.044	0.153±0.021	0.313±0.028	0.168±0.017	0.330±0.023	0.168±0.046
Song	**0.460±0.027**	0.320±0.120	0.370±0.016	0.414±0.047	0.354±0.000	0.243±0.032	0.354±0.000	0.271±0.048	0.354±0.000
WQplants	**0.173±0.004**	0.102±0.091	0.169±0.000	0.109±0.080	0.168±0.002	0.002±0.003	0.169±0.000	0.008±0.007	0.169±0.000
WQanimals	**0.096±0.005**	0.032±0.033	0.089±0.000	0.015±0.010	0.089±0.000	0.004±0.004	0.089±0.000	0.001±0.002	0.089±0.000
WaterQuality	**0.028±0.000**	0.000±0.000	**0.028±0.000**	0.000±0.000	**0.028±0.000**	0.000±0.000	**0.028±0.000**	0.001±0.002	0.021±0.014
BeLaE	**0.025±0.006**	0.008±0.004	0.017±0.005	0.002±0.002	0.016±0.006	0.010±0.006	0.016±0.005	0.011±0.004	0.016±0.005
Voice	**0.800±0.038**	0.405±0.022	0.056±0.047	0.315±0.026	0.003±0.008	0.434±0.017	0.107±0.033	0.385±0.099	0.000±0.000
Rf1	**0.092±0.039**	0.005±0.003	0.000±0.000	0.002±0.000	0.000±0.000	0.005±0.004	0.002±0.003	0.002±0.000	0.000±0.000
Thyroid	**0.799±0.007**	0.725±0.000	0.725±0.000	0.725±0.001	0.725±0.000	0.405±0.027	0.728±0.002	0.663±0.089	0.725±0.000
Pain	0.796±0.001	**0.797±0.000**	**0.797±0.000**	0.794±0.004	**0.797±0.000**	0.350±0.040	**0.797±0.000**	0.536±0.068	0.637±0.356
CoIL2000	**0.589±0.013**	0.342±0.011	0.130±0.009	0.285±0.016	0.070±0.000	0.270±0.008	0.105±0.003	0.147±0.034	0.070±0.000
TIC2000	**0.712±0.012**	0.507±0.012	0.443±0.022	0.516±0.009	0.312±0.000	0.399±0.023	0.353±0.004	0.340±0.030	0.312±0.000
Fera	**0.227±0.004**	0.186±0.000	0.186±0.000	0.169±0.017	0.186±0.000	0.021±0.005	0.189±0.001	0.033±0.014	0.186±0.000
Adult	**0.228±0.008**	0.186±0.017	0.116±0.001	0.201±0.006	0.115±0.000	0.177±0.016	0.117±0.001	0.093±0.023	0.115±0.000
Default	**0.176±0.002**	0.041±0.001	0.115±0.000	0.056±0.002	0.115±0.000	0.023±0.000	0.041±0.000	0.115±0.002	0.115±0.002
Avg. Rank	**1.323**	4.765	4.676	5.147	5.441	6.618	4.559	6.441	6.029

Table S-VIII: Test Sub-Exact Match (SEM) under 70% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	**0.969±0.000**	**0.969±0.000**	**0.969±0.000**	0.963±0.008	**0.969±0.000**	0.892±0.011	**0.969±0.000**	0.938±0.036	**0.969±0.000**
Jura	**0.846±0.037**	0.705±0.080	0.724±0.096	0.611±0.123	0.659±0.137	0.727±0.036	0.692±0.083	0.735±0.045	0.673±0.094
Enb	**1.000±0.000**	**1.000±0.000**	0.812±0.124	**1.000±0.000**	0.661±0.021	**1.000±0.000**	0.695±0.027	**1.000±0.000**	0.679±0.042
Song	**0.822±0.021**	0.803±0.083	0.725±0.020	0.820±0.020	0.709±0.000	0.808±0.019	0.710±0.003	0.799±0.034	0.709±0.000
WQplants	0.398±0.011	0.263±0.191	**0.404±0.000**	0.326±0.107	**0.404±0.000**	0.053±0.022	0.403±0.002	0.065±0.026	**0.404±0.000**
WQanimals	**0.279±0.012**	0.180±0.094	0.276±0.000	0.132±0.089	0.276±0.000	0.022±0.022	0.275±0.005	0.028±0.017	0.276±0.000
WaterQuality	**0.108±0.007**	0.018±0.024	**0.108±0.000**	0.023±0.019	**0.108±0.000**	0.000±0.000	**0.108±0.000**	0.001±0.002	0.095±0.026
BeLaE	**0.120±0.008**	0.068±0.008	0.087±0.016	0.035±0.008	0.091±0.012	0.059±0.011	0.088±0.016	0.070±0.011	0.087±0.017
Voice	**0.994±0.004**	0.942±0.011	0.689±0.030	0.896±0.016	0.582±0.033	0.934±0.016	0.691±0.027	0.935±0.020	0.553±0.000
Rf1	**0.344±0.039**	0.124±0.034	0.055±0.063	0.112±0.019	0.062±0.059	0.092±0.010	0.063±0.051	0.099±0.029	0.017±0.037
Thyroid	**0.981±0.001**	0.977±0.000	0.977±0.000	0.977±0.000	0.977±0.000	0.844±0.022	0.976±0.001	0.960±0.014	0.977±0.000
Pain	**0.895±0.001**	0.894±0.000	0.894±0.000	0.894±0.000	0.894±0.000	0.548±0.025	0.894±0.000	0.712±0.059	0.876±0.040
CoIL2000	**0.908±0.006**	0.727±0.016	0.440±0.021	0.666±0.013	0.292±0.000	0.640±0.010	0.388±0.005	0.467±0.039	0.292±0.000
TIC2000	**0.980±0.001**	0.919±0.006	0.866±0.017	0.924±0.007	0.764±0.001	0.831±0.011	0.800±0.002	0.801±0.037	0.763±0.000
Fera	**0.478±0.003**	0.326±0.000	0.326±0.000	0.326±0.010	0.326±0.000	0.162±0.009	0.334±0.002	0.163±0.027	0.326±0.000
Adult	**0.661±0.005**	0.556±0.008	0.378±0.004	0.562±0.008	0.368±0.002	0.539±0.006	0.388±0.001	0.404±0.049	0.371±0.003
Default	**0.592±0.004**	0.264±0.114	0.479±0.000	0.310±0.015	0.479±0.000	0.174±0.009	0.247±0.012	0.479±0.008	0.479±0.008
Avg. Rank	**1.588**	4.382	4.706	4.882	5.618	6.529	5.353	5.794	6.147

C.2 RUNTIME ANALYSIS.

To compare the runtime cost of the proposed algorithm, we record the average runtime of all algorithms over 5 repeated runs under each setting, as shown in the table below. We also perform Friedman tests to compute the average ranks of all comparison algorithms. Tables S-IX–S-XI show that the additional recovery-strategy optimization does not make CSER the fastest method, but its computational cost remains practical relative to the compared approaches. Its average runtime ranks are 3.812, 3.438, and 3.125 under 30%, 50%, and 70% missingness, respectively. The additional cost mainly arises from constructing repeated cross-fitted probability representations and optimizing the two interacting populations. Nevertheless, these probability representations are computed only once and kept fixed during co-evolution, avoiding repeated estimator retraining for different challenge configurations. In particular, CSER remains substantially less expensive than the ECCO-based pipelines on many medium- and large-scale datasets, although lightweight alternatives such as LEAD can be faster.

Table S-IX: Running Time under 30% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	2.037	6.894	5.488	16.304	16.110	0.129	0.051	9.232	9.537
Jura	1.934	10.486	8.923	10.861	10.644	0.033	0.083	49.901	40.489
Enb	4.646	13.507	13.295	16.260	15.847	0.132	0.178	88.764	111.129
Song	4.073	23.594	19.406	28.657	29.033	0.319	0.773	357.398	335.670
WQplants	11.219	174.703	153.798	191.557	186.414	0.592	4.818	2596.423	1821.278
WQanimals	10.414	226.580	203.423	193.698	192.909	0.823	3.305	2485.708	1513.054
WaterQuality	19.404	36.904	179.508	761.152	767.566	0.681	10.521	5110.968	3965.072
BeLaE	26.810	20.599	48.811	186.589	186.896	2.769	9.733	5752.218	3254.588
Voice	4.532	33.676	33.047	39.485	39.987	5.544	6.444	769.676	795.167
Rf1	127.256	1644.387	1607.217	1443.433	1450.128	99.337	168.662	3040.159	3147.647
Thyroid	55.541	420.672	235.715	813.147	811.386	88.412	98.582	4555.324	1401.248
Pain	686.944	900.157	557.176	1581.793	1577.737	122.766	159.447	7652.182	5654.964
CoIL2000	1408.045	877.616	611.846	1117.251	1149.746	79.381	108.771	16643.019	15287.097
TIC2000	1457.050	192.991	170.411	323.569	322.073	131.501	138.276	68690.002	61055.469
Fera	1650.995	1538.347	1280.715	1629.662	1637.774	367.154	435.030	184501.980	188786.856
Adult	1134.392	486.092	470.835	741.623	727.234	792.171	772.711	71290.535	64293.549
Default	11957.540	1504.864	1012.889	1577.988	1580.328	1041.950	856.616	185197.884	185197.884
Avg. Rank	3.812	4.688	3.938	6.062	5.938	**1.500**	2.312	8.562	8.188

Table S-X: Running Time under 50% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	1.914	6.712	4.977	16.607	16.567	0.017	0.057	23.547	19.631
Jura	1.853	10.629	8.147	10.934	10.670	0.035	0.107	50.276	39.381
Enb	1.986	14.139	13.847	16.354	15.825	0.100	0.202	79.154	97.727
Song	3.720	20.836	15.480	31.498	31.163	0.257	0.929	445.936	391.558
WQplants	9.016	215.439	144.867	193.254	195.012	0.393	6.759	2232.670	1919.985
WQanimals	8.870	195.577	135.776	187.871	193.110	0.713	5.449	1322.229	898.280
WaterQuality	16.391	37.836	265.095	758.308	755.730	0.741	20.611	3958.182	4332.277
BeLaE	22.558	18.849	51.312	187.972	187.911	0.957	14.455	5448.625	3007.560
Voice	4.472	32.781	32.191	41.082	40.664	5.360	7.712	730.525	681.897
Rf1	109.267	1572.005	1314.400	1116.291	1051.004	88.226	208.040	145328.298	139177.777
Thyroid	56.410	504.996	207.608	801.180	807.908	67.534	101.182	11041.849	2492.118
Pain	744.830	1023.684	530.638	1588.052	1582.691	134.670	191.927	62356.014	51945.215
CoIL2000	1471.972	918.028	438.173	1240.309	1224.112	118.489	176.271	70007.573	52378.252
TIC2000	1156.392	200.293	153.598	323.463	326.134	128.504	136.510	62356.014	51945.215
Fera	1289.025	2193.564	1431.363	1639.568	1641.185	367.922	469.561	145328.298	139177.777
Adult	786.370	457.678	425.708	732.886	729.073	740.097	801.835	70007.573	52378.252
Default	6825.818	1577.173	942.652	1711.592	1765.039	739.280	760.032	198628.354	198628.354
Avg. Rank	3.438	5.000	3.875	6.062	5.688	**1.375**	2.562	8.875	8.125

Table S-XI: Running Time under 70% Training Label Missingness.

Dataset	CSER	DCOM		PIST		LEAD		ECCO	
		Prototype	Correlation	Prototype	Correlation	Prototype	Correlation	Prototype	Correlation
Flare1	1.823	6.917	5.099	16.330	16.325	0.018	0.072	25.155	27.844
Jura	1.755	10.013	7.757	11.009	10.853	0.029	0.116	46.589	35.947
Enb	1.912	13.415	13.007	15.568	15.950	0.126	0.390	72.196	59.000
Song	3.665	20.868	13.941	31.118	29.461	0.314	1.370	404.764	292.131
WQplants	7.244	209.315	102.305	196.553	197.722	0.743	10.524	1703.228	722.331
WQanimals	7.255	236.619	112.947	192.993	192.515	0.624	8.297	1197.481	759.427
WaterQuality	13.800	98.742	273.388	742.150	741.489	0.689	45.839	2835.281	3513.945
BeLaE	19.118	28.925	47.784	187.006	187.103	2.137	21.274	4694.008	2306.026
Voice	3.975	32.790	29.701	39.982	41.645	5.815	8.379	707.289	451.663
Rf1	91.179	1417.126	929.782	1007.923	1023.460	96.078	265.535	133141.339	109403.256
Thyroid	46.685	500.075	194.225	771.474	762.212	81.664	125.029	14023.508	1260.621
Pain	1022.182	1140.492	401.926	1507.566	1480.823	120.137	202.982	58960.631	39589.886
CoIL2000	1370.789	899.862	332.567	1232.638	1200.611	125.745	190.828	63360.416	33804.500
TIC2000	1504.263	201.331	133.783	326.795	326.633	138.099	151.609	58960.631	39589.886
Fera	1822.838	2092.210	923.263	1619.435	1643.545	372.650	461.429	133141.339	109403.256
Adult	679.582	487.059	330.205	730.710	735.008	747.771	831.346	63360.416	33804.500
Default	4885.479	1445.073	828.998	1891.302	1904.571	741.516	760.680	208837.168	208837.168
Avg. Rank	3.125	5.062	3.562	6.000	5.875	**1.562**	2.812	8.875	8.125

D ABLATION STUDY

Our work has two main contributions: (i) the class probability representation and fusion strategy, and (ii) the evolutionary repair mechanism. To validate them, we remove each of these two components from the original CSER to construct two algorithm variants, and conduct ablation experiments on all datasets.

Ablation design. **CSER-H** removes the class-probability representation and fusion component, i.e., it does not construct or combine $\mathbf{P}^F$ and $\mathbf{P}^D$. Instead, the recovery population evolves parameters of a prototype-based rule defined from the observed labels and directly produces deterministic class assignments. When evaluating a pseudo-missing challenge, the challenged labels are excluded from prototype construction to avoid information leakage. The competitive co-evolutionary framework and the downstream MDC learning protocol are otherwise retained. This variant examines whether probabilistic class representations and their fusion provide benefits beyond deterministic hard-label recovery. **CSER-E** retains the two class-probability representations and the same confidence-aware fusion formulation, but removes competitive co-evolution and strategy selection in the final generation. Instead, the recovery strategy is determined by the corresponding fixed non-evolutionary rule. This variant evaluates the contribution of evolutionary recovery-strategy optimization while preserving the probabilistic recovery pipeline.

Evaluation protocol. The ablation study is conducted on 16 of the 17 benchmark datasets used in the main evaluation, with the same subset consistently used for all variants and missing-label rates. Tables S-XII– S-XIV report HS, EM, and SEM as mean $\pm$ standard deviation over five repeated runs with different missing-label realizations at each missing rate. Splits, observation masks, and run seeds are matched to the corresponding CSER runs. The downstream MDC learner and evaluation protocol are kept identical across all variants. CSER-H further retains the matched competitive co-evolution budget and downstream model count. Statistical comparisons are performed on the 16 paired dataset results at each missing-label rate.

Effect of probabilistic representation and fusion. Removing the class-probability representation and fusion mechanism causes a substantial degradation in downstream MDC performance. Averaged across the three missing-label rates, full CSER improves HS, EM, and SEM over CSER-H by approximately 3.9, 5.6, and 4.1 percentage points, respectively. The degradation becomes increasingly pronounced as supervision becomes sparser: the average HS gain of CSER increases from about 1.9 points at 30% missingness to 4.1 and 5.7 points at 50% and 70%, while the corresponding EM gains increase from 3.4 to 5.7 and 7.7 points. Holm-adjusted Wilcoxon signed-rank tests further show that CSER significantly outperforms CSER-H on HS, EM, and SEM at all three missing-label rates, yielding a $3/0/0$ win/tie/loss summary for every metric. For example, on Jura at 70% missingness, HS increases from 0.453 to 0.618, while EM increases from 0.223 to 0.389. These results support retaining feature-only and dependency-conditioned evidence as class-probability distributions rather than committing directly to deterministic recovered labels. Such probabilistic representations preserve predictive uncertainty and allow confidence-aware fusion to adapt their relative contributions across dimensions and instances.

Effect of evolutionary recovery optimization. Removing competitive co-evolution results in a smaller but statistically consistent performance degradation. Averaged across the three missing-label rates, full CSER improves HS, EM, and SEM over CSER-E by approximately 0.4, 0.9, and 0.4 percentage points, respectively. More importantly, the Holm-adjusted Wilcoxon signed-rank tests show that CSER significantly outperforms CSER-E on all three metrics at each of the three missing-label rates, again yielding a $3/0/0$ win/tie/loss summary for HS, EM, and SEM. These results indicate that competitive co-evolution provides a consistent additional benefit beyond the probabilistic recovery pipeline itself. Rather than relying on a fixed recovery configuration, CSER jointly evolves recovery strategies and pseudo-missing challenges, allowing candidate strategies to be optimized against diverse difficult recovery conditions. Although individual dataset–metric entries may still favor the fixed variant, the consistent cross-dataset significance supports the aggregate benefit of adaptive recovery-strategy optimization.

Implications for CSER. Taken together, the two ablations reveal complementary but asymmetric roles of the two main components of CSER. Across all nine combinations of missing rate and evaluation metric, the Friedman average ranks consistently show that CSER outperforms CSER-E, which in turn outperforms CSER-H. Removing probabilistic representation and fusion causes the larger degradation, indicating that preserving and integrating class-probability evidence provides the foundation for reliable label recovery, particularly under severe missingness. Competitive co-evolution then provides a smaller but statistically consistent additional gain by adaptively optimizing recovery strategies against diverse pseudo-missing challenges. This ordered behavior is consistent with the design of CSER: probabilistic representation and confidence-aware fusion determine how incomplete supervision is represented and integrated, whereas competitive co-evolution determines how the resulting recovery strategy is optimized. Together, these components produce recovered supervision that supports effective MDC prediction on unseen instances.

Table S-XII: Results of ablation experiments on test set metrics under 30% missingness.

Dataset	HS			EM			SEM		
	CSER	CSER-H	CSER-E	CSER	CSER-H	CSER-E	CSER	CSER-H	CSER-E
Flare1	**0.949±0.000**	0.939±0.015	0.947±0.005	**0.877±0.000**	0.862±0.022	0.874±0.007	**0.969±0.000**	0.954±0.022	0.966±0.007
Enb	**0.798±0.023**	0.737±0.009	0.795±0.025	**0.596±0.045**	0.474±0.018	0.591±0.050	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
Jura	0.666±0.014	**0.669±0.019**	0.667±0.014	**0.451±0.020**	0.439±0.029	**0.451±0.023**	0.881±0.011	**0.899±0.010**	0.886±0.015
Song	0.768±0.008	**0.778±0.007**	0.765±0.007	0.466±0.016	**0.492±0.009**	0.456±0.020	0.847±0.015	**0.852±0.015**	0.848±0.012
WQplants	**0.705±0.003**	0.703±0.003	**0.705±0.002**	0.171±0.003	0.169±0.006	**0.173±0.004**	**0.395±0.011**	0.371±0.006	0.393±0.007
WQanimals	**0.658±0.002**	0.607±0.011	0.657±0.002	**0.094±0.002**	0.023±0.011	0.091±0.004	**0.268±0.011**	0.162±0.031	0.267±0.008
WaterQuality	**0.691±0.001**	0.659±0.004	0.690±0.002	**0.028±0.000**	0.004±0.004	**0.028±0.000**	**0.115±0.004**	0.044±0.011	0.114±0.006
BeLaE	**0.432±0.003**	0.417±0.014	0.429±0.008	**0.029±0.005**	0.021±0.007	0.027±0.008	**0.140±0.005**	0.117±0.002	0.131±0.010
Voice	**0.916±0.010**	0.869±0.025	0.909±0.014	**0.833±0.019**	0.758±0.032	0.821±0.025	**0.999±0.001**	0.979±0.018	0.997±0.003
Thyroid	**0.971±0.000**	0.967±0.001	**0.971±0.001**	**0.816±0.004**	0.788±0.008	**0.816±0.005**	**0.982±0.001**	0.979±0.003	**0.982±0.001**
Rf1	**0.748±0.003**	0.709±0.012	0.741±0.012	**0.108±0.015**	0.051±0.010	0.074±0.026	**0.333±0.011**	0.278±0.022	0.318±0.031
Pain	**0.960±0.000**	**0.960±0.000**	**0.960±0.001**	**0.792±0.001**	0.788±0.000	0.788±0.005	**0.902±0.001**	0.899±0.000	**0.902±0.003**
CoIL2000	**0.932±0.002**	0.893±0.000	0.921±0.003	**0.718±0.008**	0.584±0.000	0.675±0.015	**0.945±0.003**	0.894±0.000	0.933±0.003
Fera	**0.682±0.001**	0.674±0.003	0.679±0.002	**0.247±0.002**	0.240±0.003	0.241±0.004	**0.507±0.004**	0.502±0.004	0.504±0.004
TIC2000	**0.926±0.002**	0.915±0.002	0.920±0.002	**0.791±0.003**	0.769±0.008	0.775±0.006	**0.987±0.002**	0.980±0.002	0.985±0.002
Adult	**0.710±0.001**	0.705±0.000	0.708±0.002	0.243±0.002	**0.252±0.000**	0.242±0.002	**0.669±0.003**	0.651±0.000	0.665±0.005
Avg. Rank	**1.3125**	2.6875	2.0000	**1.2812**	2.7188	2.0000	**1.3750**	2.6875	1.9375

Table S-XIII: Results of ablation experiments on test set metrics under 50% missingness.

Dataset	HS			EM			SEM		
	CSER	CSER-H	CSER-E	CSER	CSER-H	CSER-E	CSER	CSER-H	CSER-E
Flare1	**0.949±0.000**	0.936±0.004	0.948±0.002	**0.877±0.000**	0.862±0.000	0.874±0.007	**0.969±0.000**	0.946±0.011	**0.969±0.000**
Enb	**0.799±0.012**	0.701±0.014	0.780±0.009	**0.599±0.025**	0.403±0.028	0.560±0.018	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
Jura	**0.672±0.029**	0.588±0.010	0.650±0.032	**0.449±0.048**	0.324±0.000	0.424±0.053	**0.895±0.026**	0.851±0.019	0.876±0.015
Song	0.767±0.006	**0.769±0.016**	0.765±0.009	0.470±0.019	**0.477±0.038**	0.466±0.017	0.839±0.011	**0.841±0.020**	**0.841±0.016**
WQplants	**0.706±0.001**	0.688±0.018	0.704±0.002	**0.173±0.004**	0.135±0.061	0.170±0.002	**0.405±0.004**	0.342±0.040	0.398±0.008
WQanimals	**0.656±0.004**	0.556±0.025	**0.656±0.005**	0.091±0.004	0.018±0.010	**0.093±0.005**	**0.259±0.018**	0.111±0.024	0.253±0.016
WaterQuality	**0.691±0.001**	0.630±0.013	**0.691±0.001**	**0.028±0.000**	0.002±0.003	**0.028±0.000**	0.108±0.006	0.021±0.019	**0.109±0.006**
BeLaE	**0.418±0.008**	0.410±0.023	**0.418±0.009**	**0.026±0.004**	0.018±0.007	**0.026±0.008**	**0.134±0.010**	0.130±0.024	0.132±0.015
Voice	**0.911±0.004**	0.788±0.002	0.910±0.009	**0.826±0.007**	0.618±0.004	0.822±0.018	0.997±0.001	0.958±0.001	**0.998±0.002**
Thyroid	**0.971±0.001**	0.959±0.002	0.970±0.001	**0.812±0.006**	0.740±0.013	0.809±0.007	**0.982±0.001**	0.974±0.002	0.979±0.001
Rf1	**0.745±0.010**	0.666±0.012	0.740±0.015	**0.080±0.006**	0.032±0.021	0.061±0.030	**0.345±0.027**	0.162±0.038	0.325±0.048
Pain	**0.960±0.000**	0.952±0.000	0.959±0.000	**0.794±0.001**	0.749±0.000	0.792±0.002	**0.898±0.003**	0.876±0.000	**0.898±0.004**
CoIL2000	**0.918±0.002**	0.885±0.004	0.906±0.003	**0.662±0.009**	0.606±0.003	0.622±0.014	**0.932±0.006**	0.901±0.003	0.916±0.004
Fera	**0.675±0.001**	0.670±0.003	0.672±0.001	**0.241±0.000**	0.235±0.003	0.240±0.004	**0.498±0.005**	0.487±0.002	0.494±0.003
TIC2000	**0.914±0.003**	0.901±0.002	0.908±0.001	**0.759±0.010**	0.738±0.001	0.744±0.003	**0.984±0.002**	0.968±0.001	0.982±0.001
Adult	**0.707±0.002**	0.705±0.000	0.706±0.002	0.239±0.004	**0.246±0.000**	0.237±0.002	**0.662±0.005**	0.660±0.000	**0.662±0.007**
Avg. Rank	**1.1562**	2.8750	1.9688	**1.2500**	2.7500	2.0000	**1.4062**	2.8438	1.7500

Table S-XIV: Results of ablation experiments on test set metrics under 70% missingness.

Dataset	HS			EM			SEM		
	CSER	CSER-H	CSER-E	CSER	CSER-H	CSER-E	CSER	CSER-H	CSER-E
Flare1	**0.949±0.000**	0.941±0.011	**0.949±0.000**	**0.877±0.000**	0.869±0.011	**0.877±0.000**	**0.969±0.000**	0.954±0.022	**0.969±0.000**
Enb	0.758±0.044	0.680±0.025	**0.767±0.035**	0.521±0.088	0.360±0.051	**0.534±0.071**	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**
Jura	**0.618±0.024**	0.453±0.115	0.591±0.059	**0.389±0.031**	0.223±0.067	0.354±0.065	**0.846±0.037**	0.682±0.162	0.827±0.064
Song	**0.757±0.016**	0.729±0.018	0.747±0.021	**0.459±0.027**	0.367±0.042	0.434±0.040	0.822±0.021	**0.830±0.021**	0.819±0.024
WQplants	**0.706±0.002**	0.659±0.022	0.705±0.002	0.173±0.004	0.082±0.054	**0.174±0.003**	**0.398±0.011**	0.273±0.074	0.397±0.009
WQanimals	**0.662±0.002**	0.549±0.021	0.661±0.004	**0.096±0.005**	0.014±0.006	**0.096±0.007**	**0.279±0.012**	0.117±0.027	0.272±0.011
WaterQuality	**0.691±0.002**	0.591±0.023	0.687±0.005	**0.028±0.000**	0.000±0.000	0.025±0.004	**0.108±0.007**	0.006±0.002	0.099±0.015
BeLaE	**0.399±0.015**	0.396±0.020	0.391±0.013	**0.025±0.006**	0.010±0.007	0.020±0.007	**0.120±0.008**	0.093±0.025	0.105±0.011
Voice	**0.897±0.021**	0.715±0.003	0.893±0.016	**0.800±0.038**	0.507±0.011	0.795±0.029	**0.994±0.004**	0.923±0.006	0.992±0.003
Thyroid	**0.968±0.001**	0.947±0.000	**0.968±0.002**	**0.798±0.006**	0.663±0.000	0.797±0.013	**0.980±0.001**	0.965±0.000	0.979±0.001
Rf1	**0.750±0.020**	0.624±0.025	0.741±0.015	**0.092±0.039**	0.013±0.000	0.071±0.020	**0.344±0.039**	0.211±0.027	0.339±0.027
Pain	**0.959±0.000**	0.946±0.000	**0.959±0.000**	**0.796±0.001**	0.747±0.000	0.794±0.002	0.895±0.001	0.865±0.000	**0.896±0.002**
CoIL2000	**0.898±0.004**	0.890±0.005	0.887±0.002	**0.589±0.013**	0.560±0.007	0.556±0.005	**0.908±0.006**	0.900±0.010	0.893±0.011
Fera	**0.661±0.003**	0.653±0.006	0.658±0.004	**0.227±0.004**	0.221±0.005	0.224±0.003	**0.478±0.002**	0.471±0.003	0.473±0.007
TIC2000	**0.897±0.004**	0.896±0.003	0.892±0.003	**0.712±0.012**	0.699±0.005	0.696±0.009	**0.980±0.001**	0.978±0.003	0.979±0.001
Adult	**0.703±0.003**	0.694±0.000	**0.703±0.003**	0.228±0.008	**0.240±0.000**	0.229±0.010	0.661±0.005	0.639±0.000	**0.662±0.003**
Avg. Rank	**1.1875**	2.8125	2.0000	**1.3125**	2.7500	1.9375	**1.2812**	2.7500	1.9688

Table S-XV: Holm-adjusted Wilcoxon signed-rank test results for CSER and two variants under different label missing rates.

Missing	First method vs. second method	HS	EM	SEM
30%	CSER vs. CSER-H	win [3.17e-3]	win [1.23e-2]	win [9.89e-3]
	CSER vs. CSER-E	win [2.93e-3]	win [2.93e-3]	win [1.66e-2]
50%	CSER vs. CSER-H	win [1.16e-3]	win [2.24e-3]	win [3.66e-4]
	CSER vs. CSER-E	win [7.32e-4]	win [1.83e-3]	win [8.79e-3]
70%	CSER vs. CSER-H	win [1.30e-3]	win [2.33e-3]	win [1.10e-3]
	CSER vs. CSER-E	win [1.27e-2]	win [1.00e-2]	win [2.32e-3]

E SENSITIVITY TO THE EVOLUTION BUDGET.

Figure S-I compares 25, 50, and 100 generations to assess the sensitivity of CSER to the evolution budget. Across the five datasets and three missing rates, HS, EM, and SEM remain largely stable, indicating that CSER is not highly sensitive to the number of generations within the evaluated range. Increasing the budget from 25 to 50 generations often yields significant improvements, as observed on Enb, TIC2000 and WaterQuality, whereas further increasing it to 100 generations provides no consistent benefit and can slightly reduce performance in some settings. Overall, 50 generations achieves performance that is generally comparable to or better than the two alternatives while avoiding the additional cost of a longer search. We therefore adopt 50 generations as the default evolution budget.

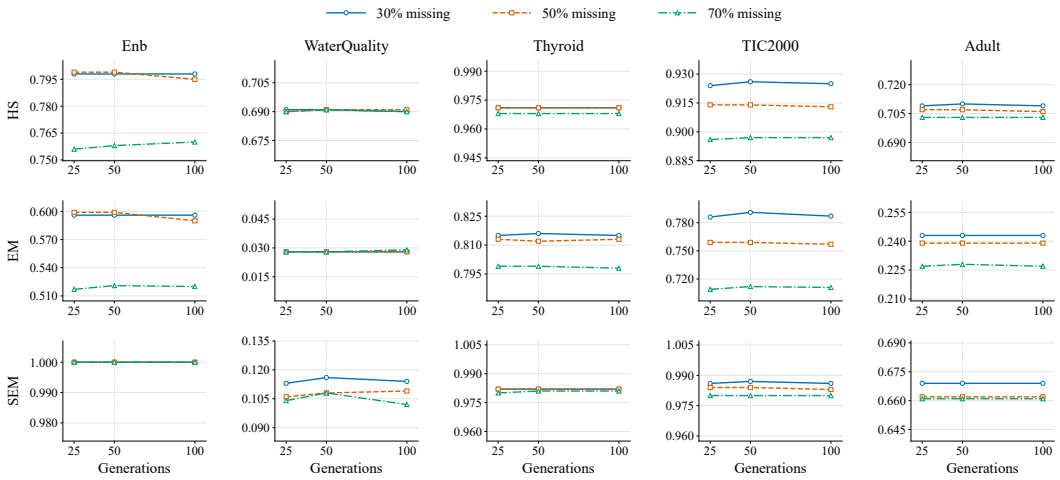


Figure S-I: Sensitivity to the evolution budget on five representative datasets. Rows correspond to HS, EM, and SEM, and columns to datasets. Curves show mean scores over five runs at 30%, 50%, and 70% missing labels. Higher values indicate better performance for all three metrics.

F Dynamic Behavior Analysis of Competitive Co-Evolution in CSER

F.1 Adaptive Evolution of the Challenge Population

To examine whether the challenge population actively adapts during competitive co-evolution, we track its generation-wise fitness dynamics on six representative datasets: Jura, WQplants, Voice, CoIL2000, TIC2000, and Adult. The same datasets are used throughout the evolutionary analysis to provide a consistent view of CSER's dynamics across different MDC problems. In particular, this analysis focuses on whether evolution progressively identifies pseudo-missing configurations that impose greater recovery difficulty on the current recovery population.

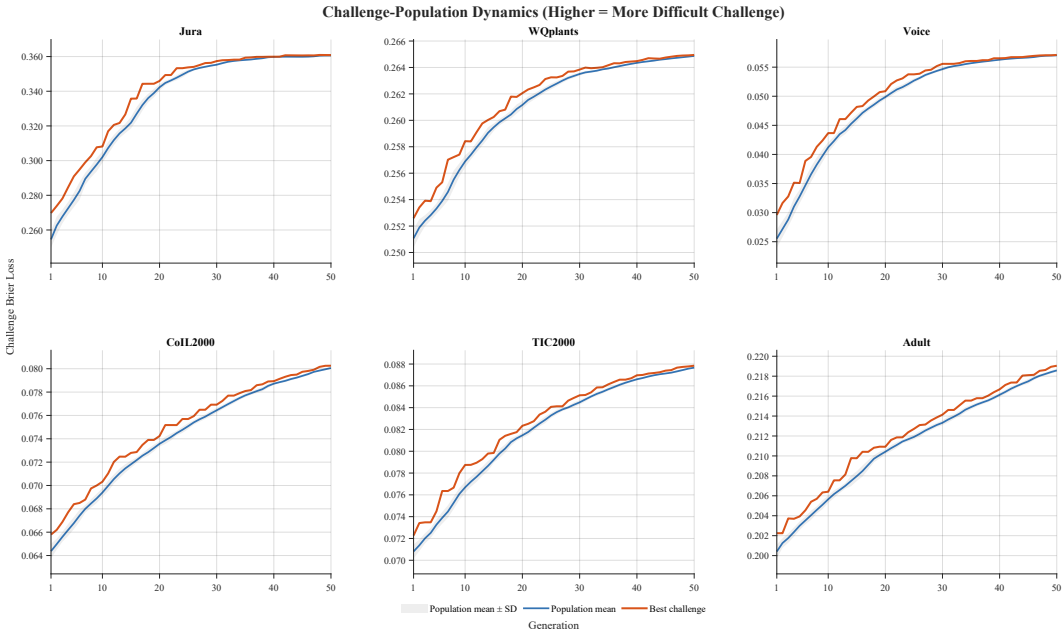


Figure S-II: Evolution of the challenge population during competitive co-evolution. The orange curve denotes the challenge with the largest mean recovery loss over the current recovery population, while the blue curve and shaded region denote the population mean and standard deviation, respectively. Higher values correspond to more difficult pseudo-missing challenges for the current recovery population.

The generation-wise challenge champion is defined as

$$b_H^{(t)} = \max_{H \in \mathcal{H}^t} F_H(H; \mathcal{R}^t). \tag{27}$$

Both the challenge champion and the population mean generally increase over generations on all six datasets. For example, the champion loss increases from 0.269 to 0.361 on Jura, from 0.029 to 0.057 on Voice, and from 0.202 to 0.219 on Adult. The stepwise increases observed especially in the early generations reflect the selection pressure toward pseudo-missing configurations that induce larger recovery errors in the current recovery population.

The evolutionary trajectories further show that the challenge search remains adaptive throughout optimization. Jura exhibits a clear slowdown after approximately generations 30–40, while WQplants and Voice also become progressively flatter. In contrast, CoIL2000, TIC2000, and Adult continue to exhibit increasing challenge fitness near the end of the evolutionary budget. These patterns indicate that the challenge population continuously adapts to the changing recovery population rather than remaining tied to a fixed missingness configuration.

Overall, the observed dynamics support the intended role of the challenge population in CSER: progressively identifying difficult pseudo-missing cases that expose weaknesses of the current recovery strategies and provide an evolving optimization environment for recovery-strategy learning.

F.2 Worst-Case Recovery Loss under Evolving Challenges

We next examine how the recovery population responds to the evolving challenge environment on the same six representative datasets. Rather than evaluating recovery strategies against a fixed challenge set, we record the largest loss faced by the generation-wise mean-fitness-optimal recovery strategy under its contemporaneous challenge population. This analysis characterizes the recovery pressure induced by the dynamically evolving challenges and complements the challenge-side dynamics above.

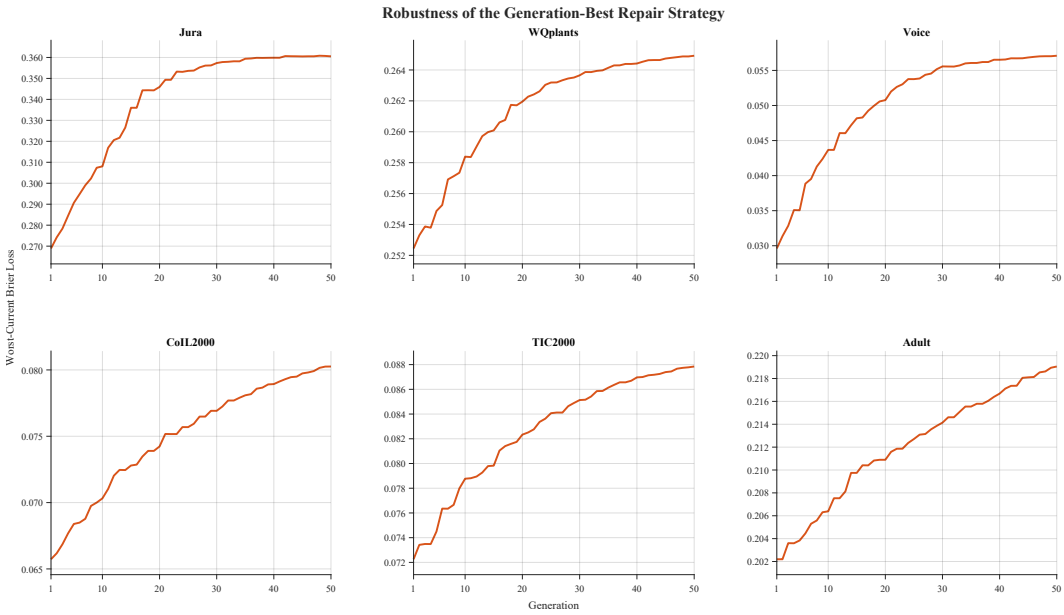


Figure S-III: Worst-case recovery loss of the generation-wise mean-fitness-optimal recovery strategy over the challenge population. The challenge set evolves together with the recovery population across generations.

For each generation, let $\boldsymbol{\theta}'_t$ denote the recovery strategy with the smallest mean loss against the current challenge population:

$$\boldsymbol{\theta}'_t \in \arg\min_{\boldsymbol{\theta}\in\mathcal{R}^t} F_R(\boldsymbol{\theta}; \mathcal{H}^t). \tag{28}$$

We further record its largest loss over the same challenge population,

$$W^{(t)} = \max_{H\in\mathcal{H}^t} L(\boldsymbol{\theta}'_t, H). \tag{29}$$

This quantity characterizes the most difficult current challenge faced by the generation-wise optimal recovery strategy.

The trajectories of $W^{(t)}$ generally increase across the six datasets, closely following the increasing difficulty of the evolving challenge population. Jura gradually approaches a plateau, while the growth on WQplants and Voice also becomes slower in later generations. CoIL2000, TIC2000, and Adult retain more visible growth throughout the 50-generation budget. This behavior is consistent with the competitive nature of CSER: while recovery strategies are repeatedly optimized against the current challenge population, the challenge population simultaneously evolves toward configurations that impose greater recovery difficulty.

Notably, at generation 50, the difference between the worst-case loss $W^{(50)}$ and the corresponding mean recovery loss is only approximately 7.71×10^{-5} to 4.72×10^{-4} across the six datasets. The small gap indicates that the mean-fitness-optimal recovery strategy exhibits relatively uniform performance over the current challenge population, rather than achieving low average loss by performing well only on a subset of easier challenges.

Together with the adaptive challenge dynamics in Fig. S-II, these results illustrate the reciprocal optimization process underlying CSER: the challenge population continually increases the difficulty of pseudo-missing recovery tasks, while the recovery population is repeatedly adapted under this evolving pressure. This interaction directly supports the motivation of CSER to learn recovery strategies under dynamically discovered missingness challenges instead of optimizing against a single fixed missingness configuration.

Overall competitive dynamics. The two analyses provide complementary views of the same co-evolutionary process. The challenge population progressively searches for harder pseudo-missing configurations, while the recovery population is continually optimized against the resulting non-stationary challenge environment. Their coupled dynamics provide empirical evidence that CSER realizes the intended interaction between missingness challenge discovery and recovery-strategy learning.