
Deciphering the Extremes: A Novel Approach for Pathological Long-tailed Recognition in Scientific Discovery

Anonymous Author(s)

Affiliation

Address

email

Abstract

Scientific discovery across diverse fields increasingly grapples with datasets exhibiting pathological long-tailed distributions: a few common phenomena overshadow a multitude of rare yet scientifically critical instances. Unlike standard benchmarks, these scientific datasets often feature extreme imbalance coupled with a modest number of classes and limited overall sample volume, rendering existing long-tailed recognition (LTR) techniques ineffective. Such methods, biased by majority classes or prone to overfitting on scarce tail data, frequently fail to identify the very instances—novel materials, rare disease biomarkers, faint astronomical signals—that drive scientific breakthroughs. This paper introduces a novel, end-to-end framework explicitly designed to address pathological long-tailed recognition in scientific contexts. Our approach synergizes a Balanced Supervised Contrastive Learning (B-SCL) mechanism, which enhances the representation of tail classes by dynamically re-weighting their contributions, with a Smooth Objective Regularization (SOR) strategy that manages the inherent tension between tail-class focus and overall classification performance. We introduce and analyze the real-world ZincFluor chemical dataset ($\mathcal{T} = 137.54$) and synthetic benchmarks with controllable extreme imbalances (CIFAR-LT variants). Extensive evaluations demonstrate our method’s superior ability to decipher these extremes. Notably, on ZincFluor, our approach achieves a Tail Top-2 accuracy of 66.84%, significantly outperforming existing techniques. On CIFAR-10-LT with an imbalance ratio of 1000 ($\mathcal{T} = 100$), our method achieves a tail-class accuracy of 38.99%, substantially leading the next best. These results underscore our framework’s potential to unlock novel insights from complex, imbalanced scientific datasets, thereby accelerating discovery. We provide the detailed code in [Appendix](#).

1 Introduction

Scientific discovery, spanning disciplines from materials science and drug development to astrophysics and genomics, increasingly relies on harnessing vast datasets. However, a pervasive and often underestimated challenge in these domains is the pathological long-tailed distribution of data. Unlike common benchmark datasets (e.g., ImageNet-LT [16], Places365-LT [22]), scientific datasets often exhibit extreme imbalances: a few well-understood or easily observable phenomena constitute the majority classes, while a multitude of rare, novel, or hard-to-characterize instances form an extensive tail. More critically, while many existing highly imbalanced benchmarks feature a large number of classes and a relatively substantial total sample size, the pathological long-tailed distributions encountered in scientific exploration are frequently characterized by a comparatively smaller number of classes coupled with a limited overall sample volume. This scarcity of available information for

each tail class imposes even more stringent demands on a model’s learning capabilities. This is not an artifact but an intrinsic feature of scientific exploration: groundbreaking discoveries often reside in these sparse tail regions, representing new materials with unique properties, biomarkers for rare diseases, or faint astronomical signals indicative of new physical laws. The criticality of accurately identifying and understanding these tail-class instances in scientific domains cannot be overstated.

Standard deep learning models and existing Long-Tailed Recognition (LTR) techniques [21, 20] often falter with such pathological imbalances (illustrated in Figure 1a or a if using subfigures). Current LTR methods, whether based on re-sampling [3, 7], re-weighting [6, 2], decoupled training [10], or specific loss designs [15], primarily aim to mitigate head-class dominance. However, with extreme scarcity, re-weighting can overfit to noise, re-sampling may lose or redundantly add information, and decoupled training struggles if initial features for tail classes are poorly learned. These shortcomings are drastically amplified at pathological imbalance levels, leading to CATASTROPHIC FAILURES in identifying scientifically paramount tail instances. For example, in our ZincFluor dataset ($T = 137.54$), rare, valuable fluorescent compounds are often missed, hindering discovery.

This paper directly confronts pathological long-tailed recognition in scientific data. We argue that extreme imbalance necessitates a *paradigm shift* from adapting existing LTR methods to designing bespoke solutions. To this end, we propose a novel, end-to-end trainable framework (overviewed in Figure 1b, with key contributions highlighted below:

- **We profoundly unveil and quantify the unique severity of the “pathological long-tail” problem within scientific discovery contexts.** By introducing and analyzing the real-world ZincFluor chemical dataset ($T = 137.54$), and complementing it with synthetic datasets we constructed featuring controllable extreme imbalance (variants of CIFAR-10-LT and CIFAR-100-LT [13]), we systematically benchmark the performance bottlenecks of existing LTR methods in these extreme scenarios, thereby providing new benchmarks and challenges for research in this domain.
- **We introduce an innovative balanced supervised contrastive learning framework, inspired by [12], engineered to fundamentally enhance the model’s capacity to perceive and represent rare yet critical scientific signals.** Our approach dynamically adjusts the contribution weights of samples from different classes during contrastive learning and integrates multi-objective optimization strategies. This not only compels the model to focus on and learn fine-grained, discriminative features for tail classes but also, through artful loss function design, ensures stable learning of common head-class phenomena. Consequently, it achieves a balanced cognitive understanding across varying class frequencies, effectively preventing the neglect of scarce signals.
- **We demonstrate the remarkable efficacy of our method through extensive evaluations.** Critically, on the highly challenging real-world ZincFluor dataset, our approach achieves a breakthrough in identifying rare fluorescent compounds, evidenced by, for instance, a Tail Top-2 accuracy of 66.84%, significantly outperforming existing techniques. Furthermore, on synthetic long-tailed benchmarks with tunable pathological imbalance, our model consistently surpasses state-of-the-art LTR methods, especially when the imbalance is more extreme. For instance, with an imbalance ratio of 1000 on CIFAR-10-LT ($T = 100$), our method achieves a tail-class accuracy of 38.99%, substantially leading the next best method at 28.55%. These results underscore the immense potential of our approach to unlock novel insights from complex, imbalanced scientific datasets, offering a potent tool to accelerate scientific discovery.

By developing a robust solution tailored to the pathological long-tailed distributions inherent in scientific research, this work aims to bridge the gap between advanced machine learning capabilities and the pressing need to extract knowledge from the most challenging, yet often most valuable, segments of scientific data.

2 Related Work

2.1 Long-Tailed Phenomena in Scientific Tasks

Long-tailed distributions, where a few common observations dominate numerous rare ones, are intrinsic to many scientific domains. For instance, in **materials science**, novel materials with exceptional functionalities are far rarer than common stable compounds [1, 17]. Similarly, **drug discovery and genomics** face challenges in identifying rare genetic variants or novel drug targets

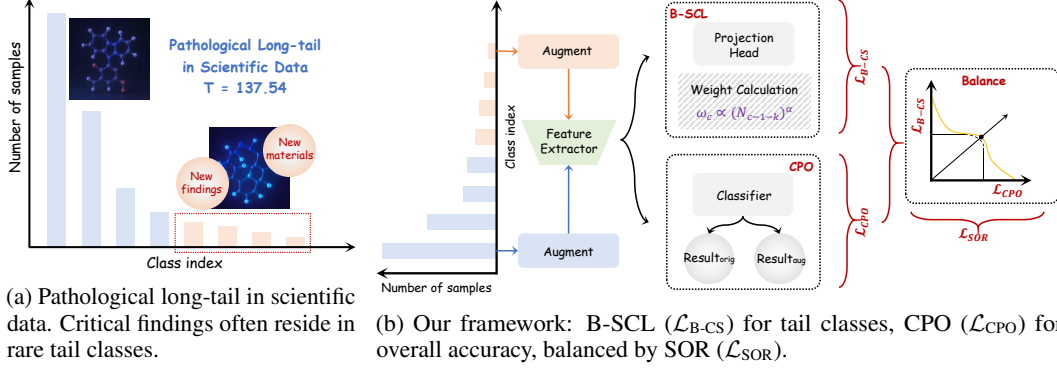


Figure 1: Visualizing (a) the pathological long-tail challenge in scientific discovery (e.g., $T = 137.54$ in the ZincFluor dataset), where critical findings are in sparse tails, and (b) our proposed framework leveraging Balanced Supervised Contrastive Learning (B-SCL), Classification Performance Objective (CPO), and Smooth Objective Regularization (SOR) to address it.

from vast datasets [4, 18]. **Astrophysics** also encounters this, with rare celestial events or objects being crucial yet sparsely observed compared to common ones [11, 8]. Distinct from typical large-scale LTR benchmarks like ImageNet-LT [16] or Places365-LT [22], scientific datasets often exhibit a pathological long-tail: extreme imbalance ratios coupled with a modest number of total classes and often limited overall sample sizes. This unique setting challenges generic LTR methods and motivates our tailored approach.

2.2 Long-Tailed Learning (LTR)

LTR techniques aim to mitigate biases towards majority classes. Broadly, these include:

- **Re-sampling strategies** balance data distribution by over-sampling minority classes (e.g., SMOTE [3]) or under-sampling majority classes [7]. However, these can lead to overfitting or information loss.
- **Re-weighting strategies** modify the loss function to assign higher importance to tail classes, examples being class-balanced loss [6], focal loss [15], and LDAM loss [2]. Careful calibration is needed to avoid issues with extremely scarce samples.
- **Decoupled learning** [10] separates representation learning from classifier training, often re-training the classifier on a balanced set. The efficacy depends heavily on the initial representation quality.
- **Other approaches** like transfer learning and knowledge distillation [9] have also been applied to LTR.

Contrastive learning for LTR is an emerging direction. Supervised Contrastive Learning (SupCon) [12] provides a strong basis for learning discriminative embeddings. Adaptations for LTR include balanced sampling or re-weighting contrastive losses [5, 14]. Our Balanced Supervised Contrastive Learning (B-SCL) specifically integrates a class-frequency aware re-weighting into the SupCon objective to handle pathological imbalances.

While most LTR methods are validated on benchmarks with many classes and samples (e.g., iNaturalist [19]), our work focuses on the distinct pathological long-tails in scientific discovery (extreme imbalance, modest class count, limited data). This necessitates a robust solution like our B-SCL with Smooth Objective Regularization (SOR) to balance learning from scarce, high-value tail data while maintaining overall performance.

3 Methodology: Balanced Contrastive Representation Learning under Dynamic Multi-Objective Constraints for Pathological Long-Tails

Our methodology addresses the critical challenge of pathological long-tailed recognition, prevalent in scientific discovery, by architecting a synergistic learning framework. This framework prioritizes the discriminative representation of tail classes while ensuring overall classification efficacy and robustness. We formalize this as a multi-objective optimization problem and derive a tractable loss function that dynamically balances these, often conflicting, objectives.

3.1 Formalizing Pathological Long-Tailed Recognition as a Multi-Objective Optimization Problem

We consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ characterized by a pathological long-tailed distribution across C classes, where $x_i \in \mathcal{X}$ and $y_i \in \{0, \dots, C-1\}$. The per-class sample count N_c exhibits extreme imbalance, quantified by $T = (\max_c N_c) / ((\min_c N_c) \cdot C)$. Our goal is to learn model parameters θ for a feature extractor f_{backbone} , a projection head π_{proj} , and a classifier g_{cls} .

In this setting, we identify three primary, potentially conflicting, learning objectives:

1. Robust Classification Performance ($\mathcal{O}_1(\theta)$): The model must achieve high classification accuracy across all classes, for both original and augmented data views. This is quantified by the Classification Performance Objective (CPO):

$$\mathcal{L}_{\text{CPO}}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell_{\text{CE}}(g_{\text{cls}}(f_{\text{backbone}}(x; \theta)), y) + \ell_{\text{CE}}(g_{\text{cls}}(f_{\text{backbone}}(x'; \theta)), y)] \quad (1)$$

where $\ell_{\text{CE}}(\mathbf{o}, y) = -\log(\text{softmax}(\mathbf{o})_y)$ is the standard cross-entropy loss. Let $\mathcal{L}_{\text{CE,orig}}(\theta) = \mathbb{E} [\ell_{\text{CE}}(g_{\text{cls}}(f_{\text{backbone}}(x; \theta)), y)]$ and $\mathcal{L}_{\text{CE,aug}}(\theta) = \mathbb{E} [\ell_{\text{CE}}(g_{\text{cls}}(f_{\text{backbone}}(x'; \theta)), y)]$. Thus, $\mathcal{L}_{\text{CPO}}(\theta) = \mathcal{L}_{\text{CE,orig}}(\theta) + \mathcal{L}_{\text{CE,aug}}(\theta)$.

2. Tail-Centric Discriminative Representation ($\mathcal{O}_2(\theta)$): The model must learn highly discriminative features, particularly for information-starved tail classes, to enable their identification. This is addressed by the Balanced Supervised Contrastive Learning (B-SCL) objective:

$$\mathcal{L}_{\text{B-SC}}(\theta) = \lambda_{\text{B-SC}} \cdot \frac{1}{2B} \sum_{\mathbf{z}_j \in \mathcal{S}_{\text{batch}}} w_{y_j} \ell_{\text{SC}}(\mathbf{z}_j; \theta) \quad (2)$$

where $\ell_{\text{SC}}(\mathbf{z}_j; \theta)$ is the standard per-anchor SupCon loss for anchor \mathbf{z}_j with label y_j , computed using embeddings $\mathbf{z} = \pi_{\text{proj}}(f_{\text{backbone}}(\cdot; \theta))$. The weights $w_c = \exp(s'_c) / \sum_k \exp(s'_k)$ with $s'_k = (N_{C-1-k})^\alpha$ up-weight tail-class contributions.

The challenge is that minimizing \mathcal{L}_{CPO} (often dominated by head classes) can conflict with minimizing $\mathcal{L}_{\text{B-SC}}$ (emphasizing tail classes). We seek a solution θ^* that is Pareto-optimal with respect to $(\mathcal{L}_{\text{CE,orig}}, \mathcal{L}_{\text{CE,aug}}, \mathcal{L}_{\text{B-SC}})$.

Optimization Target 1 (Constrained Multi-Objective Formulation) We aim to find parameters θ^* that minimize a primary combined objective while ensuring no individual sub-objective becomes excessively large. This can be conceptualized as:

$$\begin{aligned} \min_{\theta} \quad & \mathcal{L}_{\text{CPO}}(\theta) + \mathcal{L}_{\text{B-SC}}(\theta) \\ \text{subject to} \quad & \mathcal{L}_{\text{CE,orig}}(\theta) \leq \epsilon_1 \\ & \mathcal{L}_{\text{CE,aug}}(\theta) \leq \epsilon_2 \\ & \mathcal{L}_{\text{B-SC}}(\theta) \leq \epsilon_3 \end{aligned} \quad (3)$$

where $\epsilon_1, \epsilon_2, \epsilon_3$ are dynamically adjusted upper bounds.

Solving Optimization Target 1 directly is intractable. Instead, we formulate a penalty-based approach.

3.2 Derivation of the Training Objective from Multi-Objective Constraints

To find a solution approximating the Pareto front of $(\mathcal{L}_{\text{CE,orig}}, \mathcal{L}_{\text{CE,aug}}, \mathcal{L}_{\text{B-SC}})$, we employ a scalarization technique that incorporates a penalty for deviations from a balanced state.

154 **Proposition 1 (LogSumExp as a Smooth Maximum)** *The LogSumExp (LSE) function, $\text{LSE}(\mathbf{v}) =$*
 155 *$\log \sum_i \exp(v_i)$, is a differentiable, convex approximation of the maximum function, i.e., $\max_i v_i \leq$*
 156 *$\text{LSE}(\mathbf{v}) \leq \max_i v_i + \log M$ for a vector \mathbf{v} of M components.*

157 We introduce a Smooth Objective Regularization (SOR) term designed to penalize solutions where
 158 any of the fundamental objectives ($\mathcal{L}_{\text{CE,orig}}$, $\mathcal{L}_{\text{CE,aug}}$, or $\mathcal{L}_{\text{B-SC}}$) becomes disproportionately large. This
 159 aligns with the Tchebycheff (min-max) approach for multi-objective optimization. Let $\mathcal{L}_{\text{constituent}}(\theta) =$
 160 $[\mathcal{L}_{\text{CE,orig}}(\theta), \mathcal{L}_{\text{CE,aug}}(\theta), \mathcal{L}_{\text{B-SC}}(\theta)]^T$. The SOR term is defined as:

$$\mathcal{L}_{\text{SOR}}(\theta) = \lambda_{\text{SOR}} \cdot \text{LSE}(\mathcal{L}_{\text{constituent}}(\theta) / \tau_{\text{SOR}}) \quad (4)$$

161 where λ_{SOR} is a regularization strength and τ_{SOR} is a temperature parameter. For simplicity and
 162 alignment with the paper’s practical implementation, we set $\tau_{\text{SOR}} = 1$. Thus,

$$\mathcal{L}_{\text{SOR}}(\theta) = \lambda_{\text{SOR}} \cdot \log (\exp(\mathcal{L}_{\text{CE,orig}}(\theta)) + \exp(\mathcal{L}_{\text{CE,aug}}(\theta)) + \exp(\mathcal{L}_{\text{B-SC}}(\theta))). \quad (5)$$

163 The final training objective $\mathcal{L}_{\text{total}}(\theta)$ combines the primary objectives with this dynamic regularization:

$$\mathcal{L}_{\text{total}}(\theta) = \underbrace{\mathcal{L}_{\text{CE,orig}}(\theta) + \mathcal{L}_{\text{CE,aug}}(\theta)}_{\mathcal{L}_{\text{CPO}}(\theta)} + \mathcal{L}_{\text{B-SC}}(\theta) + \mathcal{L}_{\text{SOR}}(\theta). \quad (6)$$

164 Substituting Eq. 5 into Eq. 6:

$$\begin{aligned} \mathcal{L}_{\text{total}}(\theta) &= \mathcal{L}_{\text{CPO}}(\theta) + \mathcal{L}_{\text{B-SC}}(\theta) \\ &\quad + \lambda_{\text{SOR}} \cdot \log (\exp(\mathcal{L}_{\text{CE,orig}}(\theta)) + \exp(\mathcal{L}_{\text{CE,aug}}(\theta)) + \exp(\mathcal{L}_{\text{B-SC}}(\theta))). \end{aligned} \quad (7)$$

165 **Theoretical Justification.** Minimizing $\mathcal{L}_{\text{total}}(\theta)$ aims to achieve a state where: 1. The sum of the
 166 primary objectives ($\mathcal{L}_{\text{CPO}} + \mathcal{L}_{\text{B-SC}}$) is low. 2. The SOR term, leveraging Proposition 1, ensures that
 167 the maximum of the constituent objectives ($\mathcal{L}_{\text{CE,orig}}$, $\mathcal{L}_{\text{CE,aug}}$, $\mathcal{L}_{\text{B-SC}}$) is also kept low.

168 This formulation implicitly seeks a solution where no single objective can be significantly improved
 169 without degrading another, which is characteristic of Pareto-optimal solutions. The SOR term
 170 dynamically adjusts the pressure on each constituent objective. If, for instance, $\mathcal{L}_{\text{B-SC}}$ becomes very
 171 large (e.g., due to difficulty in representing extremely rare tail classes or overfitting), the gradient
 172 contribution from the SOR term with respect to $\mathcal{L}_{\text{B-SC}}$ will increase, effectively pushing the optimizer
 173 to reduce it. Similarly, if $\mathcal{L}_{\text{CE,orig}}$ is high (poor classification on original data), SOR will penalize this.

174 This dynamic balancing is crucial for pathological long-tails:

- 175 • **B-SCL** (\mathcal{O}_2) provides the necessary focus on tail classes by up-weighting their contribution to
 176 representation learning, fostering discriminative features despite data scarcity.
- 177 • **CPO** (\mathcal{O}_1) ensures general classification utility.
- 178 • **SOR** acts as the arbiter, preventing either the tail-class specific learning or the general clas-
 179 sification learning from excessively dominating and destabilizing the other, thus guiding the
 180 optimization towards a robust equilibrium suitable for the extreme imbalances encountered in
 181 scientific discovery. The [Appendix](#) provides more theory.

182 4 Detailed Methodology and Theoretical Analysis

183 This section provides a more in-depth technical exposition of our proposed methodology, particularly
 184 focusing on the formalization of the learning problem as a multi-objective optimization task and the
 185 theoretical underpinnings of the Smooth Objective Regularization (SOR) term.

186 4.1 Formalizing Pathological Long-Tailed Recognition as Multi-Objective Optimization

187 As outlined in Section ??, we model the training process for pathological long-tailed recognition as
 188 optimizing a set of distinct, potentially conflicting objectives. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$
 189 with a pathological long-tailed distribution across C classes, our goal is to learn parameters θ for a
 190 feature extractor $f_{\text{backbone}}(\cdot; \theta)$, a projection head $g_{\text{proj}}(\cdot; \theta)$, and a classifier $g_{\text{cls}}(\cdot; \theta)$.

191 The three primary objectives are:

- 192 1. **Robust Classification Performance (\mathcal{O}_1):** Minimize the Classification Performance Ob-
 193 jective (CPO), $\mathcal{L}_{\text{CPO}}(\theta)$, which measures the model’s ability to correctly classify instances
 194 from both original data (x) and augmented views (x'). This is the sum of standard cross-
 195 entropy losses on original and augmented data: $\mathcal{L}_{\text{CPO}}(\theta) = \mathcal{L}_{\text{CE,orig}}(\theta) + \mathcal{L}_{\text{CE,aug}}(\theta)$,
 196 where $\mathcal{L}_{\text{CE,orig}}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[-\log(\text{softmax}(g_{\text{cls}}(f_{\text{backbone}}(x; \theta)))_y)]$ and $\mathcal{L}_{\text{CE,aug}}(\theta) =$
 197 $\mathbb{E}_{(x,y) \sim \mathcal{D}}[-\log(\text{softmax}(g_{\text{cls}}(f_{\text{backbone}}(x'; \theta)))_y)]$.
- 198 2. **Tail-Centric Discriminative Representation (\mathcal{O}_2):** Minimize the Balanced Supervised
 199 Contrastive Learning (B-SCL) objective, $\mathcal{L}_{\text{B-SC}}(\theta)$. This objective encourages learning
 200 discriminative features, particularly for tail classes, by weighting the contribution of
 201 each sample’s contrastive loss according to its class frequency. As detailed in Appendix
 202 Section ??, this involves a weighted sum of per-sample SupCon losses: $\mathcal{L}_{\text{B-SC}}(\theta) =$
 203 $\lambda_{\text{B-SC}} \cdot \frac{1}{2B} \sum_{j \in S_{\text{batch}}} w_{y_j} l_{\text{sc}}(z_j; \theta)$, where $z_j = g_{\text{proj}}(f_{\text{backbone}}(x_j; \theta))$ and w_{y_j} are class-
 204 frequency aware weights.
- 205 3. **Objective Balancing and Stabilization (\mathcal{O}_3):** This objective is implicitly handled by the
 206 Smooth Objective Regularization (SOR) term, $\mathcal{L}_{\text{SOR}}(\theta)$, which aims to prevent any single
 207 constituent loss from becoming excessively large, thus maintaining a balanced learning
 208 process.

209 A naive approach would be to simply sum the first two objectives: $\mathcal{L}_{\text{naive}}(\theta) = \mathcal{L}_{\text{CPO}}(\theta) + \mathcal{L}_{\text{B-SC}}(\theta)$.
 210 However, as discussed in the main paper, these objectives can conflict. Minimizing \mathcal{L}_{CPO} (dominated
 211 by head classes) might pull features towards head-centric optima, while minimizing $\mathcal{L}_{\text{B-SC}}$ (with
 212 tail weighting) pulls towards tail-centric optima. Directly optimizing their sum doesn’t explicitly
 213 constrain the behavior of the individual components, potentially leading to unstable training or
 214 solutions where one objective is severely degraded.

215 Our approach addresses this by seeking a Pareto-optimal solution with respect to the constituent
 216 losses ($\mathcal{L}_{\text{CE,orig}}, \mathcal{L}_{\text{CE,aug}}, \mathcal{L}_{\text{B-SC}}$). Formally, we could conceptualize this as a constrained optimization
 217 problem: $\min_{\theta} (\mathcal{L}_{\text{CPO}}(\theta) + \mathcal{L}_{\text{B-SC}}(\theta))$ subject to: $\mathcal{L}_{\text{CE,orig}}(\theta) \leq \epsilon_1$, $\mathcal{L}_{\text{CE,aug}}(\theta) \leq \epsilon_2$, $\mathcal{L}_{\text{B-SC}}(\theta) \leq \epsilon_3$
 218 where $\epsilon_1, \epsilon_2, \epsilon_3$ are target bounds. However, setting and dynamically adjusting these bounds during
 219 training is non-trivial. We instead formulate a penalty-based approach inspired by scalarization
 220 methods for multi-objective optimization, specifically the Tchebycheff method, which minimizes the
 221 weighted L_{∞} norm of the objectives. The L_{∞} norm corresponds to the maximum value among the
 222 objectives.

223 4.2 Smooth Objective Regularization (SOR): Formal Properties and Derivation

224 To approximate the minimization of the maximum constituent objective (aligned with the Tchebycheff
 225 method) in a differentiable manner, we utilize the LogSumExp function. Proposition 1 (restated
 226 below) formally describes the LogSumExp function as a smooth approximation of the maximum.

227 **Proposition 1 (LogSumExp as a Smooth Maximum)** *The LogSumExp (LSE) function, $\text{LSE}(\mathbf{v}) =$
 228 $\log \sum_{i=1}^M \exp(v_i)$, is a differentiable, convex approximation of the maximum function, i.e., for any
 229 vector $\mathbf{v} = [v_1, \dots, v_M]^T$ of M components, we have $\max_i v_i \leq \text{LSE}(\mathbf{v}) \leq \max_i v_i + \log M$.*

230 *Proof of Proposition 1:* Let $v_{\max} = \max_i v_i$.

- 231 • Lower Bound: $\text{LSE}(\mathbf{v}) = \log \sum_{i=1}^M \exp(v_i)$. Since $\exp(v_i) > 0$ for all i , and at least
 232 one term $\exp(v_{\max})$ exists in the sum, we have $\sum_{i=1}^M \exp(v_i) \geq \exp(v_{\max})$. Taking the
 233 logarithm of both sides preserves the inequality: $\log \left(\sum_{i=1}^M \exp(v_i) \right) \geq \log(\exp(v_{\max})) =$
 234 v_{\max} . Thus, $\text{LSE}(\mathbf{v}) \geq \max_i v_i$.
- 235 • Upper Bound: We know that $v_i \leq v_{\max}$ for all i . Therefore, $\exp(v_i) \leq \exp(v_{\max})$ for all
 236 i . Summing over all M components: $\sum_{i=1}^M \exp(v_i) \leq \sum_{i=1}^M \exp(v_{\max}) = M \exp(v_{\max})$.
 237 Taking the logarithm of both sides: $\log \left(\sum_{i=1}^M \exp(v_i) \right) \leq \log(M \exp(v_{\max})) = \log(M) +$
 238 $\log(\exp(v_{\max})) = \log M + v_{\max}$. Thus, $\text{LSE}(\mathbf{v}) \leq \max_i v_i + \log M$.

Combining the bounds, we get $\max_i v_i \leq \text{LSE}(\mathbf{v}) \leq \max_i v_i + \log M$. The differentiability of LSE follows from the differentiability of the exponential and logarithm functions (for positive arguments) and summation. The convexity of LSE is a standard result, as it is the logarithm of a sum of convex functions.

The Smooth Objective Regularization (SOR) term (Equation 4 and 5) applies the LSE function to the vector of constituent losses, scaled by $1/\tau_{\text{SOR}}$. With $\tau_{\text{SOR}} = 1$, the SOR term directly approximates $\max(\mathcal{L}_{\text{CE,orig}}, \mathcal{L}_{\text{CE,aug}}, \mathcal{L}_{\text{B-SC}})$ plus an offset $\log M$ (where $M = 3$). Minimizing this term, scaled by λ_{SOR} , penalizes scenarios where any single constituent loss is significantly larger than the others.

4.3 The Total Objective and Gradient Dynamics

The total training objective is given by Equation 7: $\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{CPO}}(\theta) + \mathcal{L}_{\text{B-SC}}(\theta) + \lambda_{\text{SOR}} \cdot \log\left(\sum_{i=1}^3 \exp(\mathcal{L}_{\text{constituent},i}(\theta))\right)$ where $\mathcal{L}_{\text{constituent}} = [\mathcal{L}_{\text{CE,orig}}, \mathcal{L}_{\text{CE,aug}}, \mathcal{L}_{\text{B-SC}}]^T$.

Let's examine the gradient of \mathcal{L}_{SOR} with respect to one of the constituent losses, say $\mathcal{L}_{\text{CE,orig}}$: $\nabla_{\mathcal{L}_{\text{CE,orig}}} \mathcal{L}_{\text{SOR}}(\theta) = \lambda_{\text{SOR}} \cdot \frac{1}{\sum_{i=1}^3 \exp(\mathcal{L}_{\text{constituent},i}(\theta))} \cdot \exp(\mathcal{L}_{\text{CE,orig}}(\theta))$ This can be rewritten as $\lambda_{\text{SOR}} \cdot \frac{\exp(\mathcal{L}_{\text{CE,orig}}(\theta))}{\exp(\mathcal{L}_{\text{CE,orig}}(\theta)) + \exp(\mathcal{L}_{\text{CE,aug}}(\theta)) + \exp(\mathcal{L}_{\text{B-SC}}(\theta))}$.

This term acts as a weight or "pressure" on the gradient contribution from $\mathcal{L}_{\text{CE,orig}}$ within the SOR term. Notice that this weight is large when $\mathcal{L}_{\text{CE,orig}}$ is significantly larger than $\mathcal{L}_{\text{CE,aug}}$ and $\mathcal{L}_{\text{B-SC}}$. Specifically, if $\mathcal{L}_{\text{CE,orig}} \gg \mathcal{L}_{\text{CE,aug}}$ and $\mathcal{L}_{\text{CE,orig}} \gg \mathcal{L}_{\text{B-SC}}$, then $\exp(\mathcal{L}_{\text{CE,orig}})$ will dominate the sum, and the weight approaches $\lambda_{\text{SOR}} \cdot 1$. Conversely, if $\mathcal{L}_{\text{CE,orig}}$ is much smaller than the others, its weight approaches 0.

The gradient of the total loss with respect to the model parameters θ is: $\nabla_{\theta} \mathcal{L}_{\text{total}} = \nabla_{\theta} \mathcal{L}_{\text{CE,orig}} + \nabla_{\theta} \mathcal{L}_{\text{CE,aug}} + \nabla_{\theta} \mathcal{L}_{\text{B-SC}} + \lambda_{\text{SOR}} \cdot \sum_{i=1}^3 \frac{\exp(\mathcal{L}_{\text{constituent},i})}{\sum_{j=1}^3 \exp(\mathcal{L}_{\text{constituent},j})} \nabla_{\theta} \mathcal{L}_{\text{constituent},i}$

The last term shows how SOR contributes to the overall gradient. It applies a weighted sum of the gradients of the constituent losses, where the weights are proportional to $\exp(\mathcal{L}_{\text{constituent},i})$. This means that the objective function with the largest current value receives the highest weight in the SOR gradient term. This dynamic weighting mechanism effectively pulls parameters θ in the direction that reduces the currently largest loss component.

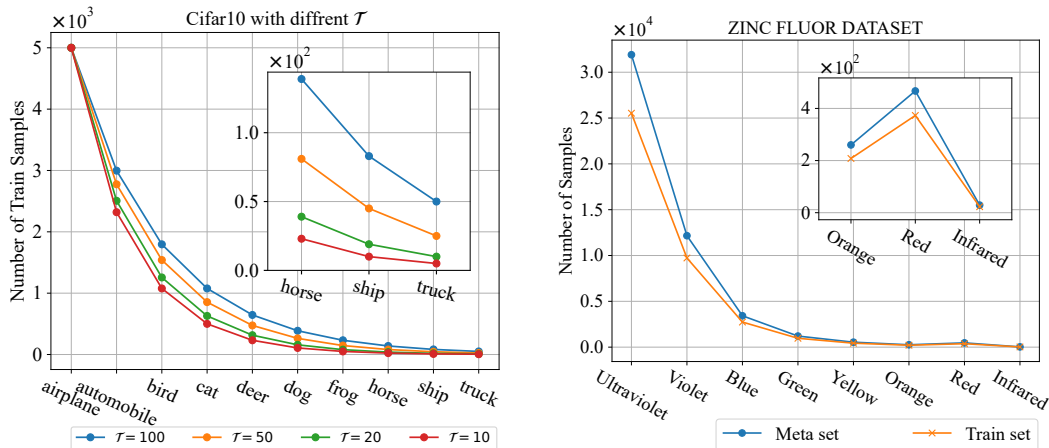
This gradient dynamic ensures that while the model optimizes the sum of the core objectives ($\mathcal{L}_{\text{CPO}} + \mathcal{L}_{\text{B-SC}}$), it is simultaneously penalized for letting any individual component loss become disproportionately large. This forces the optimization trajectory towards a region in the parameter space where all constituent losses are reasonably low, which corresponds to approximating a Pareto-optimal solution. For pathological long-tails, this prevents the optimizer from collapsing into local optima that minimize head-class loss but ignore tail classes, or vice versa, leading to a more robust and balanced model.

Summary of Mechanism:

- \mathcal{L}_{CPO} drives the model to learn general classification capabilities on both original and augmented data.
- $\mathcal{L}_{\text{B-SC}}$ focuses the representation learning on creating discriminative embeddings, specifically prioritizing tail classes via frequency-aware weighting.
- \mathcal{L}_{SOR} acts as a dynamic balancing force. By penalizing the maximum constituent loss, it prevents any single objective from dominating the learning process and ensures that all three objectives ($\mathcal{L}_{\text{CE,orig}}, \mathcal{L}_{\text{CE,aug}}, \mathcal{L}_{\text{B-SC}}$) are kept within reasonable bounds relative to each other, guiding the optimization towards a robust equilibrium necessary for deciphering pathological extremes.

5 Experiments

In this section, we conduct extensive experiments to evaluate the efficacy of our proposed method, referred to as **Ours**, in addressing pathological long-tailed recognition. We first detail the datasets



(a) Training sample distribution per class in **CIFAR-10-LT** under different \mathcal{T} settings.

(b) Comparison of meta-samples per class with training samples in **ZincFluor**.

Figure 2: Dataset characteristics: (a) **CIFAR-10-LT** class distributions. (b) **ZincFluor** sample counts.

Table 1: The anonymized **ZincFluor** dataset examples.

Index	SMILES	Pred Fluor Colour	Intensity	Fluor Value
ZINC1	<chem>CC(=O)Nc1c(-c2cccc2)c(C)nn1-c1ccc(C(=O)Nc2ccc...</chem>	Ultraviolet	Weak	1
ZINC2	<chem>Cc1nc(-c2cccc(NC(=O)c3nccn3)c2)cs1</chem>	Ultraviolet	Weak	1
ZINC3	<chem>CCCc1ccc(/N=N/C(Sc2nnc(-c3ccncc3)o2)=C(O)c2ccc...</chem>	Ultraviolet	Weak	1
ZINC4	<chem>CCOC(=O)Nc1ccc2c(Sc3cccc[n+](j3)[O-])cc(=O)oc2c1</chem>	Violet	Weak	2
ZINC5	<chem>O=CNC(=O)c1sc2ncccc3c2c1n3-c1cccc1</chem>	Violet	Weak	2
ZINC6	<chem>Cc1ccn(C(=O)c2cccc(N3CCCS3(=O)=O)c2)c=NC2CCCC...</chem>	Blue	Weak	3

and evaluation metrics (Section 5.1). We then outline the experimental setup, including baselines and implementation details (Section 5.2). Subsequently, we present quantitative results on both real-world scientific datasets and synthetic long-tailed benchmarks (Section 5.3), followed by ablation studies (Section 5.4) and qualitative analyses (Section 5.5).

5.1 Datasets, Metrics, and Pathological Imbalance

The variable \mathcal{T} is used to quantify the degree of pathological imbalance in the dataset. A higher value of \mathcal{T} corresponds to a more pronounced imbalance. It is defined as:

$$\mathcal{T} = \frac{N_{\text{majority}}}{N_{\text{minority}} \cdot N_{\text{classes}}} \quad (8)$$

where N_{majority} represents the number of samples in the majority class, N_{minority} represents the number of samples in the minority class, and N_{classes} denotes the total number of classes.

Real Dataset: ZincFluor. This is a classification dataset from a chemical laboratory. Its general content is exemplified in Table 1. As shown in Figure 2b, the dataset exhibits an extremely pathological class imbalance with an imbalance degree $\mathcal{T} = 137.54$ after an 8:2 train-test split. This severe imbalance poses a significant challenge to existing long-tailed learning methods. The dataset comprises 8 distinct fluorescence levels used as classes.

Synthetic Datasets: CIFAR-LT. To comprehensively evaluate robustness, we use long-tailed variants of **CIFAR-10** and **CIFAR-100** [13] (i.e., **CIFAR-10-LT** and **CIFAR-100-LT**). We control the imbalance ratio ($\text{IR} = N_{\text{majority}}/N_{\text{minority}}$) to construct datasets with varying degrees of pathological imbalance \mathcal{T} . Figure 2a visualizes the training sample distribution across classes in **CIFAR-10-LT** under different \mathcal{T} settings.

Evaluation Metrics. We report Top-1 accuracy as the primary metric. For **ZincFluor**, we show per-class Top-1 accuracy and aggregated tail-class accuracies (Tail Top-6, Top-4, Top-2). For **CIFAR-LT**,

we report overall Top-1 accuracy (“All”), and accuracies on “Head”, “Medium”, and “Tail” class splits based on training sample counts.

5.2 Experimental Setup

Baselines. We compare **Ours** against several long-tailed recognition baselines evaluated in prior work and relevant to our problem setting: **CE BS**, **BCL**, **CE-DRW**, **LDAM-DRW**, **KPS**, and **LORT**. For the ablation study on **ZincFluor** (Figure 3a), “base” refers to a LOS-based baseline method. For more details, please refer to [Appendix](#).

Implementation Details. All models were implemented using PyTorch and PyTorch Geometric. The experiments were conducted on a single NVIDIA Tesla A100 GPU, with results reported accordingly. Specifically, for the ZincFluor dataset, RDKit was utilized to convert SMILES strings into graph data, and a backbone network consisting of six stacked GCN layers was employed. During training, the number of epochs for the ZincFluor dataset was set to 100. For all other experiments, configurations followed those of LOS. Models were trained for 200 epochs using the SGD optimizer (learning rate $\text{lr}=0.01$, momentum=0.9, weight decay= $5\text{e-}3$) in conjunction with the CosineAnnealingLR learning rate scheduler.

5.3 Quantitative Results

Table 2: Top-1 accuracy on ZincFluor $\mathcal{T} = 137.54$. The grayed-out section indicates the primary observation indicator. **Bold** indicates the best performance while underline indicates the second best.

Method	Fluor Level								Tail Top acc		
	1	2	3	4	5	6	7	8	Top-6	Top-4	Top-2
CE	85.19	70.49	19.71	25.62	0.00	0.00	73.40	0.00	19.78	18.35	36.70
BS	82.73	30.66	43.21	28.51	0.00	25.00	72.34	0.00	28.17	24.33	36.17
BCL	86.45	51.17	51.82	22.31	17.43	40.38	69.15	50.00	<u>41.84</u>	<u>44.24</u>	59.57
CE-DRW	94.52	45.62	27.59	26.86	12.84	42.31	67.02	33.33	34.99	38.87	50.17
LDAM-DRW	91.93	47.27	28.91	20.66	22.94	28.85	69.15	33.33	33.97	38.56	51.24
KPS	91.10	45.70	51.09	23.97	1.83	19.23	71.28	0.00	27.90	23.08	35.64
LORT	72.23	25.81	1.75	33.88	0.00	26.92	75.53	0.00	23.01	25.61	37.76
Ours	90.97	42.21	58.10	21.49	11.01	34.62	67.02	66.67	43.15	44.83	66.84

Performance on ZincFluor. Table 2 details the Top-1 accuracy on **ZincFluor** ($\mathcal{T} = 137.54$). Our method demonstrates highly competitive performance on individual “Fluor Levels” and substantially outperforms all baselines in tail-class focused metrics. Notably, **Ours** achieves a Tail Top-2 accuracy of **66.84%**, a significant improvement over the second-best, **BCL** (59.57%). This underscores our method’s capability in handling real-world, pathologically imbalanced scientific data.

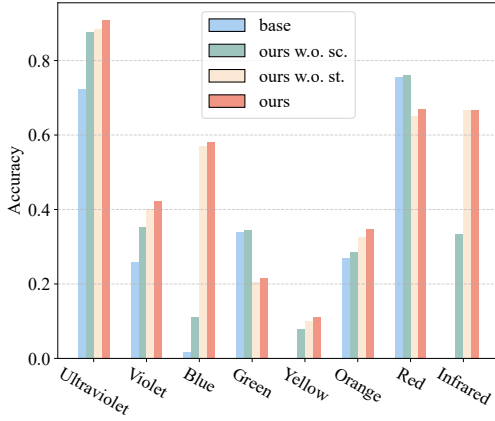
Table 3: Top-1 accuracy on CIFAR10-LT with different Imbalance ratio. The grayed-out section indicates the primary observation indicator. **Bold** indicates the best performance while underline indicates the second best.

Method	IR=1000 $\mathcal{T} = 100$				IR=500 $\mathcal{T} = 50$				IR=200 $\mathcal{T} = 20$				IR=100 $\mathcal{T} = 10$			
	Head	Medium	Tail	All	Head	Medium	Tail	All	Head	Medium	Tail	All	Head	Medium	Tail	All
CE	79.03	45.90	-	56.6	81.32	53.55	7.8	61.06	81.91	47.8	-	71.68	83.54	58.5	-	78.53
BS	76.68	64.0	16.85	62.18	76.98	69.10	30.5	66.11	82.21	61.53	-	76.01	84.81	64.8	-	80.81
BCL	79.82	57.3	28.55	<u>65.06</u>	82.22	60.05	41.25	<u>70.79</u>	82.47	71.50	-	<u>79.18</u>	83.25	81.2	-	82.84
CE-DRW	77.97	55.15	4.15	58.64	81.58	56.15	31.2	66.42	79.34	65.17	-	75.09	81.94	68.9	-	79.33
LDAM-DRW	75.57	52.0	15.25	61.19	78.27	59.75	40.7	67.05	78.79	63.7	-	74.29	81.98	68.55	-	79.29
KPS	78.9	56.85	6.65	60.04	78.95	45.2	42.75	64.96	82.27	57.23	-	74.76	82.73	61.0	-	78.38
LORT	80.75	65.30	0.05	61.52	81.0	60.0	0.05	60.61	83.36	58.50	-	75.9	83.76	85.1	-	<u>84.03</u>
Ours	76.80	76.60	38.99	69.20	81.68	79.64	59.39	77.94	84.05	84.33	-	84.14	87.59	89.80	-	88.04

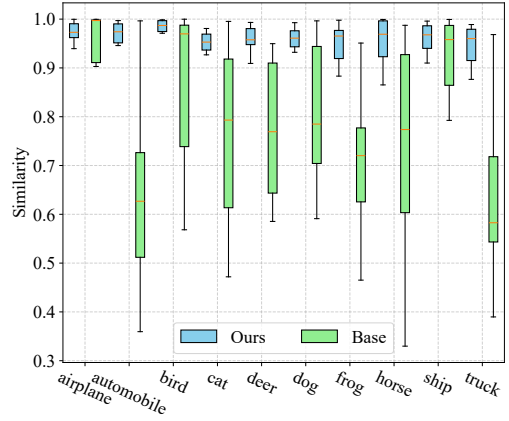
Performance on CIFAR-LT Benchmarks. Across CIFAR-LT benchmarks (Tables 3-4), our method consistently achieves superior overall accuracy and, more critically, demonstrates substantial gains in tail class accuracy across all tested imbalance ratios. For instance, on CIFAR-10-LT with extreme imbalance (IR=1000), our tail accuracy reaches **38.99%**, significantly outperforming BCL (28.55%), alongside leading overall accuracy (**69.20%** vs. 65.06%). This superior tail performance extends to CIFAR-100-LT, where at IR=100, our **32.03%** tail accuracy notably exceeds competitors (e.g., BS 27.23%), and at IR=500, we achieve **22.52%** against BCL’s 14.96%, while consistently maintaining the highest overall accuracies. These comprehensive results validate our approach’s robustness and

Table 4: Top-1 accuracy on CIFAR100-LT with different Imbalance ratio. The grayed-out section indicates the primary observation indicator. **Bold** indicates the best performance while underline indicates the second best.

Method	IR=500 $\mathcal{T} = 5$				IR=200 $\mathcal{T} = 2$				IR=100 $\mathcal{T} = 1$			
	Head	Medium	Tail	All	Head	Medium	Tail	All	Head	Medium	Tail	All
CE	80.96	46.15	7.37	36.59	79.07	51.55	6.87	42.38	78.09	48.51	10.97	47.6
BS	78.81	50.35	14.56	40.57	74.73	55.06	18.92	46.87	75.46	52.06	27.23	52.8
BCL	78.31	51.31	14.96	40.88	76.73	53.48	20.44	47.57	74.57	52.66	26.23	52.4
CE-DRW	77.58	47.08	13.58	38.93	74.87	52.55	18.71	46.05	75.89	51.69	22.07	51.27
LDAM-DRW	74.73	49.58	15.83	39.92	73.97	52.29	18.21	45.5	72.74	51.09	21.80	49.88
KPS	78.96	48.35	12.94	39.31	77.27	52.84	16.97	46.18	76.54	45.6	22.6	50.93
LORT	67.69	39.46	7.44	31.43	71.63	56.9	20.21	47.01	70.11	55.37	33.33	53.92
Ours	68.57	56.65	22.52	43.37	68.26	60.38	30.30	51.02	71.57	62.02	32.03	56.37



(a) Ablation study on **ZincFluor**. "sc." denotes B-SCL, "st." denotes SOR. Our full method outperforms ablated versions and the base.



(b) Cosine similarity between original and augmented sample features on **CIFAR-10-LT** (IR=10, trained on IR=1000). Our method shows higher robustness.

Figure 3: Ablation study and representation robustness: (a) Component analysis of our method. (b) Feature similarity across augmentations.

effectiveness in enhancing recognition of underrepresented tail classes, particularly under severe imbalance conditions.

5.4 Ablation Studies

To dissect the contributions of the core components of our method, we conduct ablation studies on the **ZincFluor** dataset, with results shown in Figure 3a. Removing the Balanced Supervised Contrastive learning loss ("sc.") from our full model ("ours") leads to a significant drop in per-class performance, particularly for the tail classes, highlighting the importance of B-SCL for learning discriminative representations under severe imbalance. Similarly, removing the Smooth Objective Regularization term ("st.") also results in degraded performance compared to the full model, indicating that SOR plays a vital role in balancing the different learning objectives and stabilizing training. The performance of our ablated models still generally surpasses the "base" LOS-based baseline. These studies confirm that both B-SCL and SOR are crucial for achieving the superior performance of our proposed framework.

5.5 Qualitative Analysis

Representation Robustness to Augmentation. Figure 3b shows the cosine similarity between the model outputs (features) of original samples and their augmented counterparts on **CIFAR-10-LT** (IR=10, models trained on IR=1000). **Ours** generally maintains higher similarity across classes compared to a **Base** method, suggesting that our approach learns representations that are more invariant and robust to data augmentations.

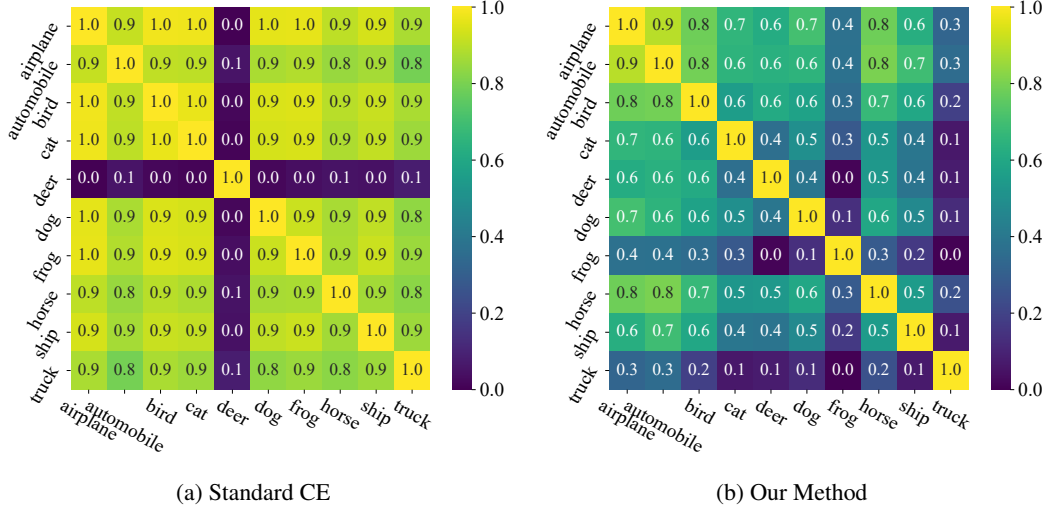


Figure 4: Class-level feature representation cosine similarities on **CIFAR-10-LT** (IR=1000). (a) Standard cross-entropy loss. (b) Our proposed method, showing improved class separability.

Class-Level Feature Discriminability. The quality of learned feature representations is further assessed by visualizing class-level cosine similarity matrices on **CIFAR-10-LT** (IR=1000), as shown in Figure 4. Panel (a) (standard **CE** loss) exhibits a diffuse similarity matrix with poor separation between classes. In contrast, panel (b) (**Ours**) displays a much clearer block-diagonal structure, indicating strong intra-class compactness and high inter-class separability. This demonstrates the superior ability of our method to learn discriminative features, which is fundamental for effective long-tailed recognition.

5.6 Discussion of Experimental Findings

The comprehensive experimental results consistently validate the efficacy of our proposed method. The substantial gains observed on the pathologically imbalanced **ZincFluor** dataset, especially in recognizing rare tail classes, highlight its practical utility for scientific discovery tasks. Furthermore, its robust and superior performance across a wide spectrum of imbalance ratios on synthetic **CIFAR-LT** benchmarks underscores its generalizability and strength in handling varying degrees of data imbalance. The ablation studies confirm the synergistic contributions of the B-SCL and SOR components, and qualitative analyses provide visual evidence of the improved representation quality and feature discriminability achieved by our approach. These findings strongly support our central claim that a tailored framework integrating balanced contrastive representation learning with dynamic multi-objective optimization is pivotal for effectively addressing pathological long-tailed recognition.

6 Conclusion

This paper tackled the critical issue of pathological long-tailed recognition in scientific discovery, where rare instances crucial for breakthroughs are often missed by standard methods. We introduced a novel framework combining Balanced Supervised Contrastive Learning (B-SCL) to enhance tail-class representation and Smooth Objective Regularization (SOR) to dynamically balance competing learning objectives. Our approach ensures focused learning on sparse tail data without compromising overall performance. Extensive experiments on the real-world ZincFluor dataset and synthetic CIFAR-LT benchmarks with extreme imbalances demonstrated significant improvements over state-of-the-art LTR techniques, particularly in identifying critical tail classes. This work provides a robust tool for extracting valuable insights from severely imbalanced scientific datasets, paving the way for accelerated discovery. Future directions include incorporating domain knowledge and extending to other scientific data modalities.

References

- [1] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018.
- [2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.
- [3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [4] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the royal society interface*, 15(141):20170387, 2018.
- [5] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 715–724, 2021.
- [6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.
- [7] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, 2003.
- [8] Christopher J Fluke and Colin Jacobs. Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1349, 2020.
- [9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [10] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.
- [11] Ajit Kembhavi and Rohan Pattnaik. Machine learning in astronomy. *Journal of Astrophysics and Astronomy*, 43(2):76, 2022.
- [12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [14] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6918–6928, 2022.
- [15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [16] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019.

- 431 [17] Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in
432 materials science with emphasis on adaptive sampling using uncertainties for targeted design.
433 npj Computational Materials, 5(1):21, 2019.
- 434 [18] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George
435 Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine
436 learning in drug discovery and development. Nature reviews Drug discovery, 18(6):463–477,
437 2019.
- 438 [19] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig
439 Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection
440 dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition,
441 pages 8769–8778, 2018.
- 442 [20] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition.
443 International Journal of Computer Vision, 130(7):1837–1872, 2022.
- 444 [21] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learn-
445 ing: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(9):10795–
446 10816, 2023.
- 447 [22] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A
448 10 million image database for scene recognition. IEEE transactions on pattern analysis and
449 machine intelligence, 40(6):1452–1464, 2017.

A ZincFluor Dataset: Molecular Fluorescence and Pathological Imbalance

This section provides a more detailed introduction to the ZincFluor dataset used in our experiments (Section 5.2). We describe its origin, the underlying scientific task of molecular fluorescence prediction, and elaborate on why its inherent data distribution constitutes a "pathological long-tail" challenge, distinct from standard long-tailed benchmarks.

A.1 Molecular Fluorescence and Its Scientific Significance

Molecular fluorescence is a photophysical process where a molecule absorbs light at a specific wavelength and then re-emits light at a longer wavelength. This phenomenon is of immense scientific and technological importance, underpinning applications in diverse fields such as:

- **Materials Science:** Development of organic light-emitting diodes (OLEDs), fluorescent probes for material characterization, and luminescent sensors.
- **Biology and Medicine:** Bioimaging (using fluorescent tags or probes), drug discovery (screening for compounds with desired fluorescence properties or using fluorescent markers), and diagnostic assays.
- **Chemistry:** Understanding molecular electronic structure, reaction monitoring, and analytical techniques.

Discovering and synthesizing novel molecules with tailor-made fluorescence properties (e.g., specific emission wavelengths, high quantum yield, photostability, environmental sensitivity) is a core pursuit in chemistry and materials science. Predicting these properties computationally from molecular structure is a valuable tool to accelerate this discovery process, avoiding costly and time-consuming experimental synthesis and screening.

A.2 The ZincFluor Dataset: Structure and Task

The ZincFluor dataset originates from experimental measurements conducted in a chemical laboratory. It comprises a collection of molecular compounds, each represented by its SMILES (Simplified Molecular Input Line Entry System) string, which provides a concise textual description of the molecule’s structure. For each compound, an associated fluorescence property has been measured and categorized into one of 8 distinct levels. These levels, serving as the classification labels in our task, represent different aspects of the molecule’s fluorescent behavior, such as intensity or emission characteristics (exemplified by the 'Pred Fluor Colour', 'Intensity', and 'Fluor Value' columns in Table 1 in the main paper). The task is to predict the fluorescence level of a molecule given its SMILES string (or its graph representation derived from it, as described in Section ??).

A.3 Pathological Imbalance in ZincFluor

As highlighted in the main paper (Figure 2b), the ZincFluor dataset exhibits an exceptionally severe class imbalance across its 8 fluorescence levels. Following an 8:2 training-test split, the imbalance factor T (defined as $N_{\text{majority}} / (N_{\text{minority}} \cdot N_{\text{classes}})$) is calculated to be 137.54. This metric quantifies the extreme disparity between the most frequent and least frequent fluorescence levels relative to the total number of classes.

The distribution is not merely skewed; it is **pathologically** imbalanced in the context of scientific discovery for the following key reasons:

1. **Scientific Value Concentrated in the Tail:** The rare fluorescence levels (minority classes) often correspond to molecules exhibiting unusual, extreme, or highly specific fluorescent properties. These are precisely the properties that are most sought after for cutting-edge applications (e.g., a molecule with exceptionally high brightness for bioimaging, or a compound emitting light at a very specific wavelength for sensing). In essence, the "rare" instances in this dataset represent the potential scientific breakthroughs or novel materials, not just less common variations of typical compounds.
2. **Limited Sample Volume:** Unlike large-scale public benchmarks like ImageNet-LT or Places365-LT which have millions of images and hundreds/thousands of classes, scientific

498 datasets like ZincFluor often arise from costly and labor-intensive experimental processes.
 499 This results in a relatively modest total sample size (approximately [Insert Total Sample
 500 Count Here] samples for ZincFluor). The combination of extreme imbalance and limited
 501 total data volume means that the tail classes have critically few samples, sometimes only a
 502 handful, making robust learning for these classes incredibly difficult.

503 3. **Fundamental Task Objective:** The implicit goal in analyzing such scientific data is often
 504 to discover or identify these rare, high-value instances for further investigation. A model
 505 that performs well on the frequent (head) classes but fails to identify the rare fluorescent
 506 compounds in the tail effectively misses the primary scientific objective.

507 The ZincFluor dataset represents a "pathological long-tail" problem because the tail instances are
 508 not just rare data points, but are scientifically **critical** signals buried within a sea of common
 509 observations. Standard LTR methods, primarily designed to mitigate majority bias in general
 510 classification tasks, struggle significantly with this unique combination of extreme imbalance, limited
 511 data, and the paramount importance of correctly identifying the scarce tail instances that drive
 512 scientific advancement. Our work is specifically motivated to address this particular challenge.

513 B Evaluation on Custom Places-LT Datasets with Reduced Class Counts

514 To further demonstrate the robustness and generalizability of our method across different domains
 515 and varying degrees of pathological imbalance complexity, we conducted additional experiments on
 516 modified versions of the standard Places-LT dataset [22]. While the original Places-LT features a
 517 large number of classes (365), our definition of pathological long-tail in scientific discovery contexts
 518 highlights scenarios with extreme imbalance coupled with a modest number of classes and limited
 519 overall sample volume(Section 1). To better align with this, we constructed synthetic long-tailed
 520 datasets derived from Places-LT that maintain a high imbalance ratio but reduce the total number of
 521 classes.

522 B.1 Custom Places-LT Dataset Construction

523 We generated new synthetic datasets based on the original Places-LT dataset (with an imbalance ratio
 524 of 996) by performing class-level sampling. Specifically, we followed a procedure to select a subset
 525 of classes:

- 526 • We identified the class with the maximum number of samples (most frequent) and the class
 527 with the minimum number of samples (least frequent) in the original Places-LT dataset.
 528 These two classes were always included in our custom datasets.
- 529 • From the remaining classes in the original Places-LT, we uniformly sampled additional
 530 classes until the desired total number of categories was reached.

531 Using this procedure, we constructed three new datasets containing a total of 10, 50, and 100 classes,
 532 respectively. These datasets are referred to as "Places-LT (IR 996, 10 Categories)", "Places-LT (IR
 533 996, 50 Categories)", and "Places-LT (IR 996, 100 Categories)".

534 Importantly, these newly constructed datasets maintain the same imbalance ratio ($IR =$
 535 $N_{majority}/N_{minority} = 996$) as the original Places-LT. However, by significantly reducing the to-
 536 tal number of classes while keeping a very high IR, these datasets exhibit an even more pronounced
 537 "pathological" nature in terms of the *relative sparsity of tail classes within a smaller overall class
 538 space*, aligning more closely with the characteristics observed in datasets like ZincFluor compared
 539 to the original 365-class Places-LT. Figure 5 shows the sample distribution characteristics for these
 540 three custom datasets, illustrating the persistent long-tail across varying numbers of categories.

541 B.2 Experimental Setup

542 B.3 Results and Analysis

543 Table 5 presents the quantitative results on the three custom Places-LT datasets.

544 The results show that our method consistently achieves the best overall Top-1 accuracy across all
 545 three variants of the custom Places-LT datasets, ranging from 10 to 100 categories, while maintaining

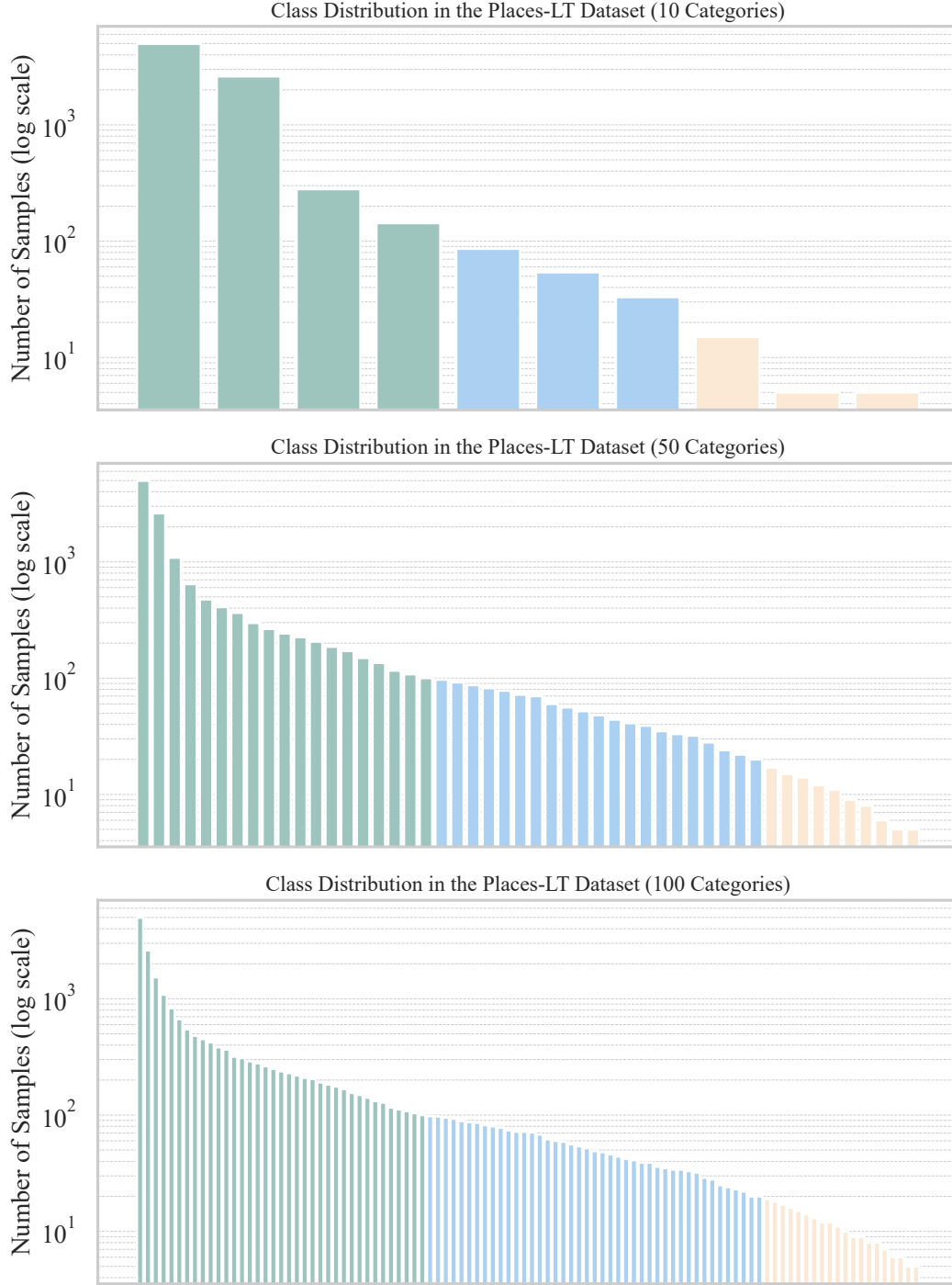


Figure 5: The Sample Distribution Characteristics of Various Datasets Generated via Places-LT (IR 996) with 10, 50, and 100 Categories.

546 a very high imbalance ratio (996). More critically, our method demonstrates superior performance on
 547 the challenging Tail classes in all scenarios. For the 10-category dataset, Ours achieves a Tail accuracy
 548 of **62.0%**, significantly outperforming the next best method, LDAM-DRW (60.33%). As the number
 549 of classes increases to 50 and 100, while maintaining the extreme IR, the Tail task becomes even more
 550 challenging, and many baselines exhibit severely degraded performance (e.g., CE, CE-DRW, KPS,
 551 LORT showing near zero or very low tail accuracy). In these more complex pathological settings

Table 5: Top-1 accuracy on Custom Places-LT Datasets (IR 996) with 10, 50, and 100 Categories. The grayed-out column indicates the overall accuracy. **Bold** indicates the best performance while underline indicates the second best.

Method	IR 996 / 10 Categories				IR 996 / 50 Categories				IR 996 / 100 Categories			
	Head	Medium	Tail	All	Head	Medium	Tail	All	Head	Medium	Tail	All
CE	61.50	0.00	0.00	24.60	40.28	0.00	0.00	14.50	28.14	0.00	0.00	10.13
BS	93.25	56.00	39.00	65.80	66.17	51.00	35.60	53.38	53.39	43.27	32.95	44.85
BCL	89.00	82.00	56.00	77.00	77.22	62.23	39.10	63.00	63.78	58.52	50.70	58.85
CE-DRW	94.50	62.67	37.67	67.90	66.28	44.36	15.10	46.44	53.97	39.77	16.90	40.31
LDAM-DRW	91.00	77.00	60.33	<u>77.60</u>	76.83	62.73	43.10	<u>63.88</u>	62.75	55.75	44.30	<u>55.97</u>
KPS	91.00	85.67	15.67	66.80	77.28	48.77	3.90	50.06	63.56	45.77	8.90	44.80
LORT	95.75	14.00	0.00	42.50	50.61	0.32	0.00	18.36	36.11	3.70	0.05	14.64
Ours	85.00	77.50	62.0	83.60	60.22	76.71	58.50	67.13	47.92	59.81	53.55	54.27

(50 and 100 categories), Ours continues to achieve the highest Tail accuracy (**58.50%** and **53.55%** respectively), substantially leading the second best methods (LDAM-DRW at 43.10% and 44.30%).

Furthermore, Ours also shows strong performance on Medium classes (best on 50 and 100 categories) and competitive, albeit not always leading, performance on Head classes (best on 100 categories), demonstrating its ability to balance learning across the entire frequency spectrum. The strong overall performance (**83.60%**, **67.13%**, **54.27%**) underscores this capability.

These experiments on custom Places-LT datasets with varied, but modest, class counts further validate the effectiveness of our proposed framework in handling pathological long-tailed distributions beyond the specific domain of ZincFluor. They highlight our method’s unique ability to maintain high accuracy on scarce tail classes while ensuring robust overall performance, a critical requirement for scientific discovery applications.

C Detailed Theoretical Analysis

This section provides a detailed theoretical analysis of our proposed framework, emphasizing its formulation as an implicitly constrained multi-objective optimization problem and the role of Smooth Objective Regularization (SOR) in achieving a dynamically balanced solution for pathological long-tailed recognition. Our approach deviates from methods solely focused on manipulating classification logits or sample/loss weights based on class frequencies, by instead directly influencing the optimization trajectory to balance competing objectives via a principled penalty mechanism.

We define the model parameters as $\theta \in \mathbb{R}^P$. The learning process is driven by three constituent loss functions defined on the dataset \mathcal{D} :

- $\mathcal{L}_1(\theta) = \mathcal{L}_{\text{CE,orig}}(\theta)$: Cross-Entropy loss on original data.
- $\mathcal{L}_2(\theta) = \mathcal{L}_{\text{CE,aug}}(\theta)$: Cross-Entropy loss on augmented data.
- $\mathcal{L}_3(\theta) = \mathcal{L}_{\text{B-SC}}(\theta)$: Balanced Supervised Contrastive Learning loss, which incorporates frequency-aware weighting w_y for tail classes (details in Appendix Section ??).

Let $\mathbf{L}(\theta) = [\mathcal{L}_1(\theta), \mathcal{L}_2(\theta), \mathcal{L}_3(\theta)]^T \in \mathbb{R}^3$ be the vector of constituent losses.

In the context of pathological long-tailed data, minimizing the simple sum $\sum_{k=1}^3 \mathcal{L}_k(\theta)$ can lead to suboptimal solutions where one loss is significantly higher than others. Our framework implicitly aims to solve a constrained multi-objective optimization problem: finding θ^* that minimizes a primary objective while keeping all constituent losses below certain thresholds ϵ_k . This is conceptually similar to:

$$\min_{\theta} \sum_{k=1}^3 \mathcal{L}_k(\theta) \quad \text{s.t.} \quad \mathcal{L}_k(\theta) \leq \epsilon_k, \quad k = 1, 2, 3.$$

Directly solving this is challenging. We approximate this via a penalty method, augmenting the sum of losses with a term that penalizes the maximum of the constituent losses.

Proposition 2 (LogSumExp as a Smooth Maximum) *The LogSumExp (LSE) function, $\text{LSE}(\mathbf{v}) = \log \sum_{i=1}^M \exp(v_i)$, is a differentiable, convex approximation of the maximum function, satisfying $\max_i v_i \leq \text{LSE}(\mathbf{v}) \leq \max_i v_i + \log M$ for any vector $\mathbf{v} = [v_1, \dots, v_M]^T \in \mathbb{R}^M$.*

587 Let $v_{\max} = \max_{i \in \{1, \dots, M\}} v_i$. For the lower bound:

$$\sum_{i=1}^M \exp(v_i) \geq \exp(v_{\max}). \quad (9)$$

588 Taking the logarithm (monotonic):

$$\log \left(\sum_{i=1}^M \exp(v_i) \right) \geq \log(\exp(v_{\max})) = v_{\max}. \quad (10)$$

589 Thus, $\max_i v_i \leq \text{LSE}(\mathbf{v})$.

590 For the upper bound: Since $v_i \leq v_{\max}$ for all i , $\exp(v_i) \leq \exp(v_{\max})$. Summing:

$$\sum_{i=1}^M \exp(v_i) \leq \sum_{i=1}^M \exp(v_{\max}) = M \exp(v_{\max}). \quad (11)$$

591 Taking the logarithm:

$$\log \left(\sum_{i=1}^M \exp(v_i) \right) \leq \log(M \exp(v_{\max})) = \log M + \log(\exp(v_{\max})) = \log M + v_{\max}. \quad (12)$$

592 Thus, $\text{LSE}(\mathbf{v}) \leq \max_i v_i + \log M$. Differentiability and convexity are standard properties of LSE.

593 We define the Smooth Objective Regularization (SOR) term using the LSE function applied to our
594 vector of constituent losses $\mathbf{L}(\theta)$:

$$\mathcal{L}_{\text{SOR}}(\theta) = \lambda_{\text{SOR}} \cdot \text{LSE}(\mathbf{L}(\theta)) / \tau_{\text{SOR}} \quad (13)$$

595 where $\lambda_{\text{SOR}} > 0$ and $\tau_{\text{SOR}} > 0$. In our implementation, we set $\tau_{\text{SOR}} = 1$ and $M = 3$, yielding:

$$\mathcal{L}_{\text{SOR}}(\theta) = \lambda_{\text{SOR}} \cdot \log(\exp(\mathcal{L}_1(\theta)) + \exp(\mathcal{L}_2(\theta)) + \exp(\mathcal{L}_3(\theta))). \quad (14)$$

596 Minimizing \mathcal{L}_{SOR} directly penalizes the largest constituent loss, pushing it down relative to the others.

597 Our total training objective is defined as the sum of the constituent losses augmented by the SOR
598 term:

$$\mathcal{L}_{\text{total}}(\theta) = \sum_{k=1}^3 \mathcal{L}_k(\theta) + \mathcal{L}_{\text{SOR}}(\theta) \quad (15)$$

599 Substituting the expression for \mathcal{L}_{SOR} :

$$\mathcal{L}_{\text{total}}(\theta) = \sum_{k=1}^3 \mathcal{L}_k(\theta) + \lambda_{\text{SOR}} \cdot \log \left(\sum_{j=1}^3 \exp(\mathcal{L}_j(\theta)) \right). \quad (16)$$

600 Minimizing $\mathcal{L}_{\text{total}}$ serves as a penalty method approximation to the constrained multi-objective
601 problem. It drives down the sum of losses while simultaneously using \mathcal{L}_{SOR} to keep the maximum
602 loss value in check, acting as a soft "tail penalty" (when \mathcal{L}_3 is high) and a "smooth constraint"
603 promoting balance across all objectives.

604 The dynamic balancing property is evident from the gradient of $\mathcal{L}_{\text{total}}(\theta)$:

$$\nabla_{\theta} \mathcal{L}_{\text{total}}(\theta) = \sum_{k=1}^3 \nabla_{\theta} \mathcal{L}_k(\theta) + \nabla_{\theta} \mathcal{L}_{\text{SOR}}(\theta). \quad (17)$$

605 The gradient of the SOR term is derived using the chain rule. Let $L_k = \mathcal{L}_k(\theta)$:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{SOR}}(\theta) &= \lambda_{\text{SOR}} \cdot \nabla_{\theta} \log \left(\sum_{j=1}^3 \exp(L_j) \right) \\ &= \lambda_{\text{SOR}} \sum_{k=1}^3 \frac{\partial \log(\sum_{j=1}^3 \exp(L_j))}{\partial L_k} \nabla_{\theta} L_k \\ &= \lambda_{\text{SOR}} \sum_{k=1}^3 \frac{\exp(L_k)}{\sum_{j=1}^3 \exp(L_j)} \nabla_{\theta} \mathcal{L}_k(\theta). \end{aligned}$$

606 Let $p_k(\theta) = \frac{\exp(\mathcal{L}_k(\theta))}{\sum_{j=1}^3 \exp(\mathcal{L}_j(\theta))}$. These p_k values form a probability distribution over the constituent
 607 losses, where p_k is high when \mathcal{L}_k is large. The total gradient becomes:

$$\nabla_{\theta} \mathcal{L}_{\text{total}}(\theta) = \sum_{k=1}^3 \nabla_{\theta} \mathcal{L}_k(\theta) + \lambda_{\text{SOR}} \sum_{k=1}^3 p_k(\theta) \nabla_{\theta} \mathcal{L}_k(\theta). \quad (18)$$

608 Rearranging terms:

$$\nabla_{\theta} \mathcal{L}_{\text{total}}(\theta) = \sum_{k=1}^3 (1 + \lambda_{\text{SOR}} p_k(\theta)) \nabla_{\theta} \mathcal{L}_k(\theta). \quad (19)$$

609 Equation 19 reveals the dynamic balancing. The gradient of each constituent loss $\nabla_{\theta} \mathcal{L}_k$ contributes
 610 to the total gradient with a weight $(1 + \lambda_{\text{SOR}} p_k)$. When a specific loss \mathcal{L}_k becomes significantly larger
 611 than others, $p_k \rightarrow 1$, and the weight $(1 + \lambda_{\text{SOR}} p_k) \rightarrow 1 + \lambda_{\text{SOR}}$, effectively amplifying the gradient
 612 $\nabla_{\theta} \mathcal{L}_k$. Conversely, for a small loss \mathcal{L}_j , $p_j \rightarrow 0$, and its gradient is weighted by approximately 1.
 613 This mechanism ensures that the optimization actively targets the largest loss component, pulling it
 614 down.

615 This adaptive weighting, governed by the relative magnitudes of $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3$, imposes the "smooth
 616 constraint" by discouraging any single loss from dominating. For pathological long-tails, this is
 617 crucial: it prevents the model from solely optimizing the easily satisfied CPO on head classes while
 618 neglecting the critical \mathcal{L}_3 for tail classes, and vice-versa. Instead, it promotes a balanced decrease
 619 across all objectives, leading to a more robust model capable of deciphering the challenging extremes.
 620 This principled approach, rooted in approximating constrained multi-objective optimization via
 621 SOR's gradient dynamics, provides a theoretical basis for our method's superior performance on
 622 pathological long-tailed data.