# 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:

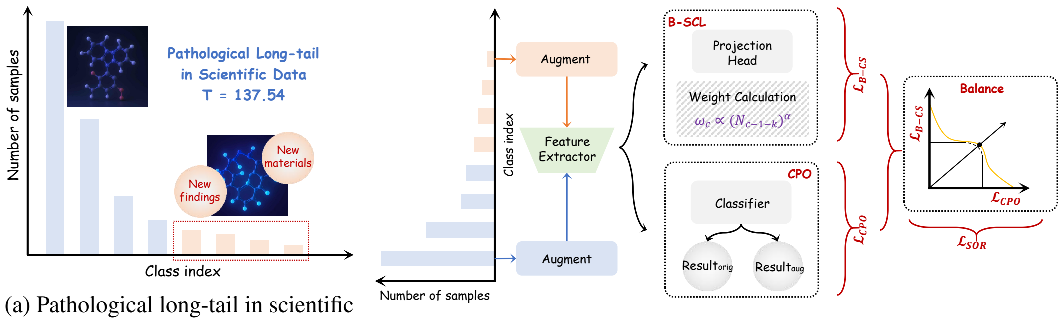
- **We profoundly unveil and quantify the unique severity of the "pathological long-tail" problem within scientific discovery contexts.** By introducing and analyzing the real-world ZincFluor chemical dataset ($T = 137.54$), and complementing it with synthetic datasets we constructed featuring controllable 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




(a) Pathological long-tail in scientific data. Critical findings often reside in rare tail classes.

(b) Our framework: B-SCL ($\mathcal{L}_{\text{B-CS}}$) for tail classes, CPO ($\mathcal{L}_{\text{CPO}}$) for overall accuracy, balanced by SOR ($\mathcal{L}_{\text{SOR}}$).

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 $\theta$ for a feature extractor $f_{\text{backbone}}$, a projection head $\pi_{\text{proj}}$, and a classifier $g_{\text{cls}}$.

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):

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

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

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

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 $\mathcal{L}_{\text{CPO}}$ (often dominated by head classes) can conflict with minimizing $\mathcal{L}_{\text{B-SC}}$ (emphasizing tail classes). We seek a solution $\theta^*$ 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 $\theta^*$ that minimize a primary combined objective while ensuring no individual sub-objective becomes excessively large. This can be conceptualized as:*

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

*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.

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

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

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

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

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

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

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

Substituting Eq. 5 into Eq. 6:

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

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

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

This dynamic balancing is crucial for pathological long-tails:

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

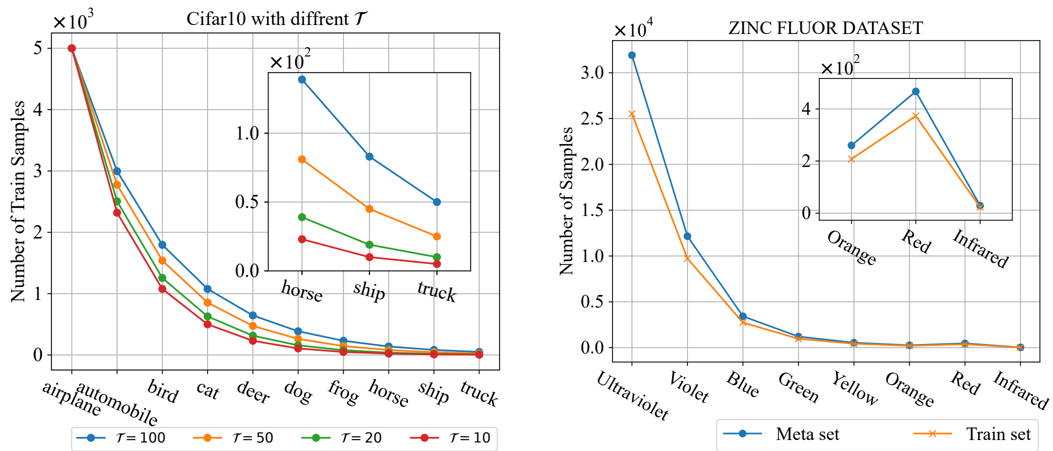
# 4 Experiments

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 and evaluation metrics (Section 4.1). We then outline the experimental setup, including baselines and implementation details (Section 4.2). Subsequently, we present quantitative results on both real-world scientific datasets and synthetic long-tailed benchmarks (Section 4.3), followed by ablation studies (Section 4.4) and qualitative analyses (Section 4.5).

## 4.1 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:

$$\mathcal{T} = \frac{N_{\text{majority}}}{N_{\text{minority}} \cdot N_{\text{classes}}} \tag{8}$$




(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 | CC(=O)Nc1c(-c2cccccc2)c(C)nn1-c1ccc(C(=O)Nc2ccc... | Ultraviolet | Weak | 1 |
| ZINC2 | Cc1nc(-c2cccc(NC(=O)c3ncccn3)c2)cs1 | Ultraviolet | Weak | 1 |
| ZINC3 | CCCc1ccc(/N=N/C(Sc2nnc(-c3ccncc3)o2)=C(O)c2ccc... | Ultraviolet | Weak | 1 |
| ZINC4 | CCOC(=O)Nc1ccc2c(Sc3ccccc[n+]3[O-])cc(=O)oc2c1 | Violet | Weak | 2 |
| ZINC5 | O=CNC(=O)c1sc2ncccc3c2c1ncn3-c1cccccc1 | Violet | Weak | 2 |
| ZINC6 | Cc1ccn(C(=O)c2cccc(N3CCCS3(=O)=O)c2)c=NC2CCCC... | Blue | Weak | 3 |

where $N_{\text{majority}}$ represents the number of samples in the majority class, $N_{\text{minority}}$ represents the number of samples in the minority class, and $N_{\text{classes}}$ 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 (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.

### 4.2 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 lr=0.01, momentum=0.9, weight decay=5e-3) in conjunction with the CosineAnnealingLR learning rate scheduler.

### 4.3 Quantitative Results

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

| Method | Fluor Leval | | | | | | | | 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 | 41.84 | 44.24 | 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 $\mathbf{66.84}\%$, a significant improvement over the second-best, **BCL** ($59.57\%$). This underscores our method's capability in handling real-world, pathologically imbalanced scientific data.

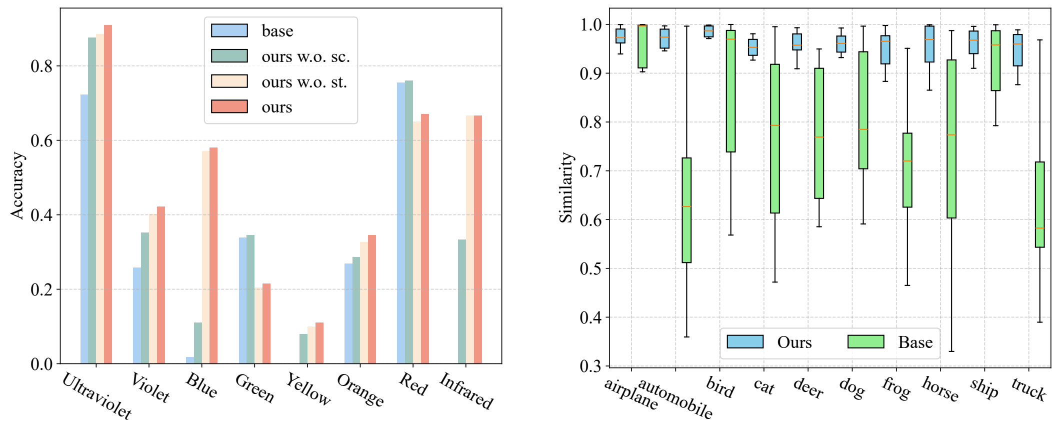
Table 3: Top-1 accuracy on CIFAR10-LT with different Imbalance ratio. The grayed-out section indicates the primary observation indicator. **Blod** indicates the best performance while underline indicates the second best.

| 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 | 65.06 | 82.22 | 60.05 | 41.25 | 70.79 | 82.47 | 71.50 | - | 79.18 | 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 | - | 84.03 |
| 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** |

Table 4: Top-1 accuracy on CIFAR100-LT with different Imbalance ratio. The grayed-out section indicates the primary observation indicator. **Blod** 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** |

**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 $\mathbf{38.99}\%$, significantly outperforming BCL ($28.55\%$), alongside leading overall accuracy ($\mathbf{69.20}\%$ vs. $65.06\%$). This superior tail performance extends to CIFAR-100-LT, where at IR=100, our $\mathbf{32.03}\%$ tail accuracy notably exceeds competitors (e.g., BS $27.23\%$), and at IR=500, we achieve $\mathbf{22.52}\%$ against BCL's $14.96\%$, while consistently maintaining




(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.

the highest overall accuracies. These comprehensive results validate our approach's robustness and effectiveness in enhancing recognition of underrepresented tail classes, particularly under severe imbalance conditions.

### 4.4 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.

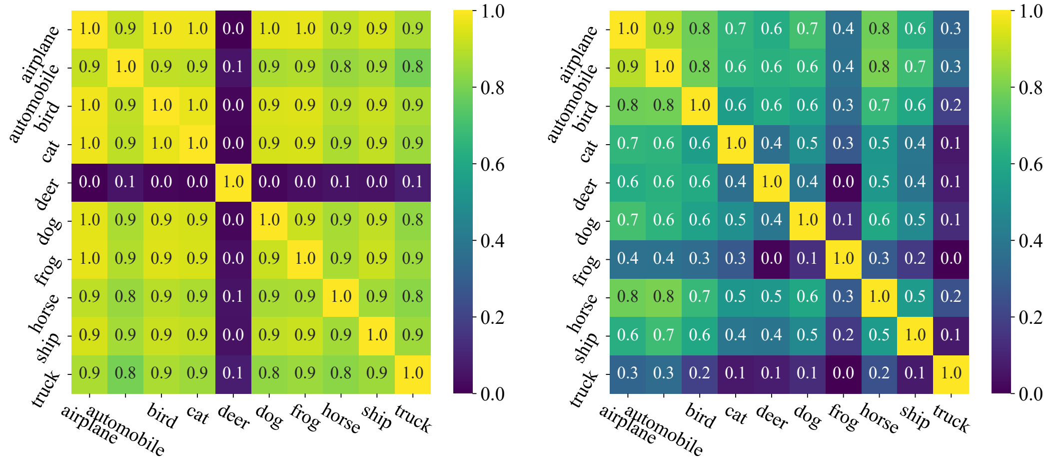
### 4.5 Qualitative Analysis

**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.

**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.

### 4.6 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




(a) Standard CE

(b) Our Method

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.

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.

# 5 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.

[17] Turab Lookman, Prasanna V Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Computational Materials, 5(1):21, 2019.

[18] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. Nature reviews Drug discovery, 18(6):463–477, 2019.

[19] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8769–8778, 2018.

[20] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. International Journal of Computer Vision, 130(7):1837–1872, 2022.

[21] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(9):10795–10816, 2023.

[22] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence, 40(6):1452–1464, 2017.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We list our contributions in detail at the end of the Introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the relevant assumptions and proofs in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code links in the abstract and implementation details in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the source of the dataset and code in Appendix.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We describe the relevant details in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report performance metrics such as average accuracy on multiple datasets.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mention the computational resources used for the experiments in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read and understood the code of ethics; and have done our best to conform.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our experiments are conducted on publicly available data. It does no impact the society at large, beyond improving our understanding of certain aspects of deep learning and crystal struction prediction.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We are not shipping our code with any source code or binary files from any other existing libraries, so there are no concerns over getting permission or including a license. We did cite open-sourced libraries, e.g. PyTorch, in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code base with included readme files. We do not ship any source code or binary files from any other existing libraries.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our work does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.