

STELLAR: Spatio-Temporal Environmental Learning with Latent Alignment and Refinement for Long-Tailed Species Distribution Modeling

Shufeng Kong^{1,2}, Tao Yu¹, Yuanyuan Wei¹, Caihua Liu^{5,3,2*}, Junwen Bai², Yingheng Wang², Marc Grimson², Daniel Fink⁴, Carla P. Gomes²

¹ School of Software Engineering, Sun Yat-sen University, Zhuhai, China

² Department of Computer Science, Cornell University, Ithaca, NY, USA

³ Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA

⁴ Cornell Lab of Ornithology, Ithaca, NY, USA

⁵ School of Computer Science and Artificial Intelligence, Foshan University, Foshan, China
{sk2299, cl2869, jb2467, yw2349, mg2425, daniel.fink}@cornell.edu, gomes@cs.cornell.edu,
{yutao67, weiyi53}@mail2.sysu.edu.cn

Abstract

Joint Species Distribution Modeling (JSDM) is a key enabler for biodiversity monitoring and conservation planning. However, accurate JSDM faces two coupled challenges: environmental drivers and species distributions are inherently spatio-temporal, while species co-occurrence patterns exhibit complex non-linear community structure and severe long-tail imbalance driven by rare species. Existing approaches often address these factors in isolation, learning from static covariates or neglecting the historical trajectories of dynamic community structure. To overcome these limitations, we propose STELLAR (Spatio-Temporal Environmental Learning with Latent Alignment and Refinement), a novel framework that learns a shared latent space where dynamic habitat context and community structure are optimized jointly. Our approach integrates three complementary components: (1) a Graph-Temporal Encoder that employs graph attention and recurrent units to aggregate spatial neighborhood effects and capture the co-evolving historical dynamics of environmental context and community structure; (2) a Context-Anchored Latent Alignment mechanism that structures the latent space using a label-activated mixture prior and supervised contrastive learning, actively clustering species based on shared environmental preferences; and (3) an Imbalance-Aware Decoupled Decoding module that utilizes Asymmetric Loss to focus learning on hard, rare species samples, preventing mode collapse in the long tail. Experiments on the large-scale eBird dataset, curated with domain experts, demonstrate that our framework significantly outperforms state-of-the-art baselines, particularly in predicting rare species and revealing interpretable species interactions.

1 Introduction

The accelerating biodiversity crisis, exacerbated by anthropogenic climate change and habitat fragmentation, necessitates scalable computational paradigms to monitor ecosystems, anticipate range shifts, and prioritize conservation interventions [Gomes *et al.*, 2019]. While Species Distribution Models (SDMs) [Elith and Leathwick, 2009] serve as the cornerstone of this effort, linking occurrences to environmental covariates, ecological communities are not merely aggregates of independent species. Real-world assemblages are governed by two tightly coupled forces: (i) *extrinsic spatio-temporal dynamics*, where habitat suitability is shaped by accumulating environmental conditions and historical trajectories, and (ii) *intrinsic biotic structure*, which induces complex, non-linear co-occurrence regimes and severe long-tail imbalance, where the majority of species are rare but ecologically critical [Wisz *et al.*, 2013; Ovaskainen and Soininen, 2011]. Despite advances in deep learning for ecology, current Joint Species Distribution Models (JSDMs) often confront a fundamental mismatch between this ecological reality and modeling practice. Here we identify three critical gaps in the existing literature:

The Spatio-Temporal Representation Gap. On the *input side*, standard pipelines typically treat location-time samples as independent, static snapshots. This assumption violates two key ecological regularities: *spatial spillover* (source-sink dynamics) and *temporal persistence* (ecological memory). While statistical methods like Gaussian Processes (GPs) [Golding and Purse, 2016] and recent Spatial Implicit Neural Representations (SINR) [Cole *et al.*, 2023] improve spatial continuity, they generally assume isotropic smoothing. They fail to model *anisotropic barriers* (e.g., rivers blocking dispersal) and rarely capture the causal neighborhood graph explicitly. Furthermore, autoregressive models used in weather forecasting [Fan *et al.*, 2022] have yet to be effectively adapted to the irregular, sparse nature of biodiversity survey data, leaving the co-evolution of environments and assemblages under-modeled.

*Correspondence to Caihua Liu: cl2869@cornell.edu

The Structural Alignment Gap. On the *output side*, community prediction requires learning high-dimensional label structures. Classical statistical approaches, such as Hierarchical Modeling of Species Communities (HMSC) [Ovaskainen *et al.*, 2017] and Deep Multivariate Probit (DMVP) [Chen *et al.*, 2018], rely on Multivariate Probit (MVP) formulations. While statistically rigorous, MVP models scale cubically with species richness $\mathcal{O}(S^3)$, rendering them intractable for large-scale species monitoring. Deep latent-variable approaches, such as Multivariate Probit Variational AutoEncoder (MPVAE) [Bai *et al.*, 2020], offer scalable alternatives. However, these models typically rely on restrictive unimodal Gaussian priors ($\mathcal{N}(0, I)$). In heterogeneous ecosystems, this isotropic assumption forces distinct community types (e.g., forest guilds vs. grassland guilds) to collapse into a single mode. While Contrastive Gaussian Mixture VAEs (C-GMVAE) [Bai *et al.*, 2022] have attempted to introduce multimodal priors, they rely on static feature encoders, failing to align the latent community structure with the dynamic trajectory of the environmental context and species assemblages.

The Long-Tail Inference Gap. Finally, a critical failure mode in JSMD is the handling of rare species. Biodiversity data follows a heavy-tailed distribution; even in species-rich areas, most species are absent in any given survey. Standard reconstruction objectives, such as Binary Cross Entropy (BCE), are dominated by these abundant “easy negatives.” While computer vision has successfully addressed imbalance via loss re-weighting strategies like Focal Loss [Lin *et al.*, 2017] or Logit Adjustment [Menon *et al.*, 2021], these advances have not been widely adopted in generative ecological modeling. Furthermore, traditional data resampling techniques are effectively impossible in multi-label settings, as a single survey frequently contains both head (common) and tail (rare) species. Consequently, generative JSMDs, such as C-GMVAE and MPVAE, frequently collapse into a trivial degenerate solution: predicting “absent” for all rare species, rendering the model ineffective for conservation planning.

Stakeholder Collaboration. This work was conducted in direct collaboration with domain experts and conservation biologists from the Cornell Lab of Ornithology. Their roles included: (1) **Problem Formulation:** Our partners identified the failure of current models to predict rare species (“the long tail”) as a primary bottleneck for conservation planning, directly motivating our Imbalance-Aware design; (2) **Data Curation:** Domain experts facilitated access to quality-controlled detection pipelines and advised on the selection of ecologically relevant environmental covariates; and (3) **Validation:** Biologists qualitatively evaluated the ecological validity of our model predictions and latent structure. This ensures the proposed framework meets the operational standards required for biodiversity monitoring.

Our Contribution. In this work, we address these challenges by reformulating JSMD as a joint representation learning and robust structural alignment problem. We propose a novel framework that synthesizes the co-evolving trajectories of environmental context and species assemblages with an imbalance-aware generative model. Our framework integrates three tightly coupled components:

First, we learn spatio-temporal habitat embeddings via a hybrid Graph-Temporal Encoder. By integrating Graph Neural Networks (GNN) with Recurrent Neural Networks (RNN), this module (i) propagates information across anisotropic spatial neighborhoods to capture dispersal effects, and (ii) summarizes the historical trajectories of both environmental covariates and antecedent community states. This process yields a context-rich, dynamic embedding that encapsulates the co-evolution of space and time, serving as a critical conditioning signal for downstream community modeling.

Second, we adopt a Context-Anchored Latent Alignment mechanism to resolve multimodal community structures. We formulate a label-activated mixture prior utilizing learnable species prototypes as distinct cluster centroids. By leveraging supervised contrastive learning, this mechanism aligns the spatio-temporal habitat embedding with these centroids, creating a dynamic “gravitational field” in the latent space: the environmental representation of a site is actively pulled toward the prototypes of its co-occurring species, ensuring the learned context captures the correct ecological regime.

Third, to robustly model the “long tail” of biodiversity, we introduce an Imbalance-Aware Decoupled Decoding strategy. By integrating Asymmetric Loss (ASL) [Ridnik *et al.*, 2021] directly into the variational objective, we replace standard reconstruction with a dynamic focusing mechanism. This induces a “hard-mining” effect during generation, systematically down-weighting the gradients from abundant easy negatives to redirect model capacity toward the rare, difficult-to-predict species that drive conservation priorities.

2 Methodology

Accurate biodiversity forecasting requires resolving two coupled challenges: modeling the *extrinsic spatio-temporal dynamics* of the environment, where drivers propagate across locations and accumulate over history, and capturing the *intrinsic multimodal structure* of species assemblages, which exhibit severe class imbalance. We propose a novel framework comprising three components: (1) a **GNN-RNN Encoder** for context-rich biophysical embedding; (2) a **Context-Anchored Latent Alignment** mechanism for multimodal structure learning; and (3) an **Imbalance-Aware Decoupled Decoding** module for robust long-tail prediction. Figure 1 illustrates the architecture of our model.

2.1 Preliminaries and Problem Formulation

Let $\mathcal{S} = \{(u, t, \mathbf{x}_{u,t}, \mathbf{y}_{u,t})\}$ denote a set of observed survey records, where u represents a specific geographic location and t is a discrete time step. For each record, we observe a local environmental covariate vector $\mathbf{x}_{u,t} \in \mathbb{R}^D$ (e.g., land cover) and a multi-label species presence-absence vector $\mathbf{y}_{u,t} \in \{0, 1\}^S$, where S is the species richness.

To explicitly model regional connectivity and historical dynamics despite irregular sampling, we discretize the spatial domain into a grid system $\mathcal{C} = \{c_1, \dots, c_M\}$. We construct a dynamic spatial graph $\mathcal{G}_t = (\mathcal{C}, \mathcal{E}_t)$, where nodes $c \in \mathcal{C}$ represent grid cells and edges $e \in \mathcal{E}_t$ encode spatial adjacency or functional connectivity between cells.

We define the *regional context* for a grid cell c as a temporal window of size W containing aggregated covariates and

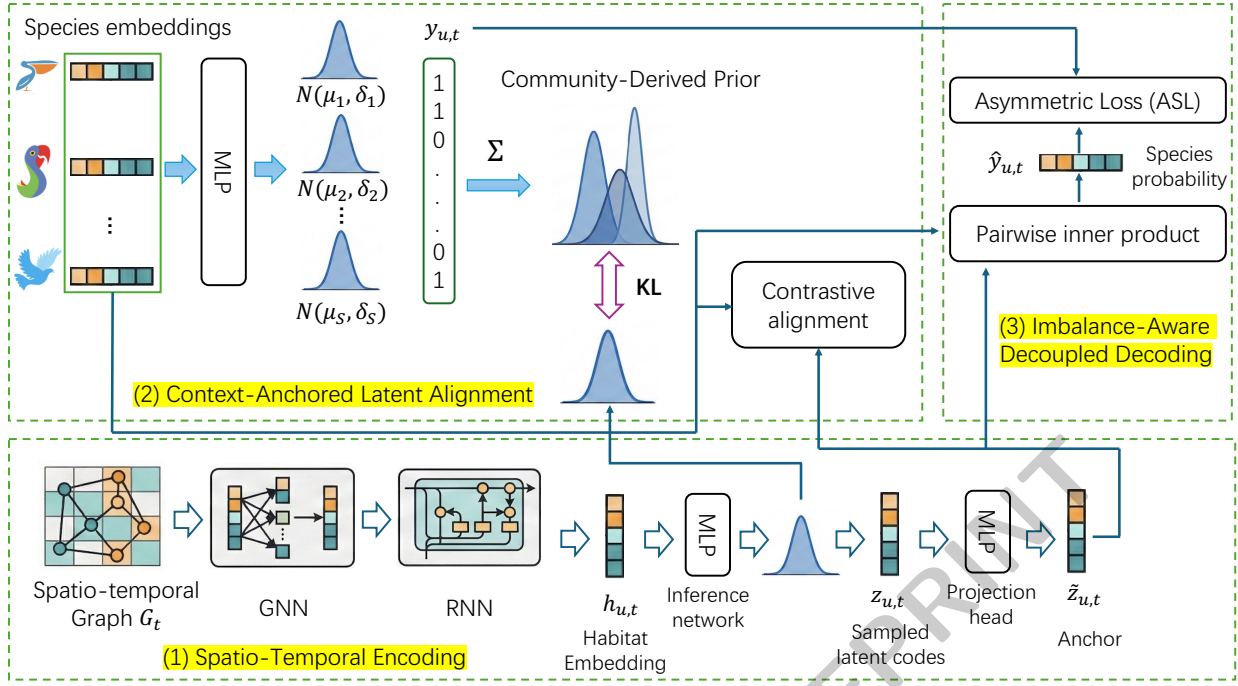


Figure 1: The architecture of our model STELLAR: (1) A Spatio-Temporal Encoder (GNN+RNN) fuses regional and historical cues into a habitat embedding $h_{u,t}$. (2) The Context-Anchored Latent Alignment module aligns the variational posterior with a label-activated Gaussian Mixture Prior using KL divergence and contrastive learning. Note that while the distributions are depicted as univariate contours for visual clarity, they are modeled as high-dimensional multivariate Gaussians in the latent space. (3) An Imbalance-Aware Decoupled Decoding module utilizes ASL to focus learning on hard, rare species samples, preventing mode collapse in the long tail.

community states: $\mathbf{H}_{c,t}^{\text{grid}} = [(\bar{\mathbf{x}}_{c,\tau}, \bar{\mathbf{y}}_{c,\tau})]_{\tau=t-W}^{t-1}$, where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are pooled averages of all surveys within cell c at time τ .

Our objective is to learn a probabilistic model $P_{\Theta}(\mathbf{y}_{u,t} | \mathbf{x}_{u,t}, \mathbf{H}_{c(u),t}^{\text{grid}}, \mathcal{G})$ that maximizes the likelihood of the observed assemblages, conditioned on both the high-resolution local environment $\mathbf{x}_{u,t}$ and the coarse-grained regional history $\mathbf{H}_{c(u),t}^{\text{grid}}$ of the enclosing cell $c(u)$.

2.2 Spatio-Temporal Habitat Representation

Ecological dynamics are governed by processes that are both spatially diffusive (e.g., dispersion) and temporally cumulative (e.g., extinction debts). Standard Convolutional Neural Networks (CNNs) assume a fixed grid structure, which is often mismatched with the irregular distribution of ecological survey data. To capture these dynamics while retaining local precision, we introduce a hierarchical **Point-Grid Fusion strategy** backed by a serialized GNN-RNN architecture.

Spatio-Temporal Grid Aggregation. Ecological survey data is frequently unstructured and opportunistic (e.g., citizen science checklists), leading to irregular spatial sampling. To regularize this input, we discretize the spatial domain into a grid system. For each time step t , we map all available survey checklists to their corresponding grid cells.

For a grid cell c , we compute an aggregated feature vector $\mathbf{x}_{c,t}$ by pooling the environmental covariates (averaged within the cell) and the species presence-absence records (ag-

gregated from all checklists in the cell). This aggregation mitigates sampling noise and aligns disparate data sources into a structured representation.

Grid-Based Graph Construction. We construct the spatial graph $\mathcal{G}_t = (\mathcal{C}, \mathcal{E}_t)$ where nodes $c \in \mathcal{C}$ represent the grid cells. Edges $e \in \mathcal{E}_t$ are defined based on spatial adjacency (connecting neighboring grid cells) or strictly defined distance thresholds between grid centroids. This topology allows the subsequent GNN layers to propagate information across the grid lattice, effectively modeling the diffusion of species and environmental drivers from neighboring regions into the focal cell.

It is worth noting that while our current grid adjacency approach captures dominant spatial gradients, the framework is extensible. Future iterations could define edges using functional connectivity matrices (e.g., following wind or ocean currents) to explicitly model vector-based dispersal without altering the underlying architecture.

Spatial Propagation (Regional Context). Raw environmental measurements at a single location are often insufficient due to the Mass Effect paradigm, where species populations are sustained by the stability of the surrounding meta-community. However, spatial influence is rarely uniform; barriers (e.g., rivers, elevation changes) can decouple geographically close regions.

To model these non-uniform dependencies, we employ a

Graph Attention Network (GAT) [Veličković *et al.*, 2017] on the grid graph. We define the combined node feature vector at time τ as $\mathbf{f}_{c,\tau} = [\mathbf{x}_{c,\tau} \parallel \mathbf{y}_{c,\tau}]$, concatenating current covariates with the community state. At each time step τ , we compute the attention coefficients for every edge (c, v) :

$$e_{cv}^\tau = \text{LeakyReLU}(\mathbf{a}^\top [\mathbf{W}_s \mathbf{f}_{c,\tau} \parallel \mathbf{W}_s \mathbf{f}_{v,\tau}]), \quad (1)$$

where $\mathbf{W}_s$ is a learnable projection matrix, $\mathbf{a}$ is the attention vector, and $\parallel$ denotes concatenation. We normalize these coefficients using a softmax function to obtain the final attention weights α_{cv}^τ :

$$\alpha_{cv}^\tau = \frac{\exp(e_{cv}^\tau)}{\sum_{k \in \mathcal{N}(c) \cup \{c\}} \exp(e_{ck}^\tau)}. \quad (2)$$

The regionally contextualized feature vector $\mathbf{s}_{c,\tau}$ is then computed as a weighted sum of the neighbors:

$$\mathbf{s}_{c,\tau} = \sigma \left(\sum_{v \in \mathcal{N}(c) \cup \{c\}} \alpha_{cv}^\tau \mathbf{W}_s \mathbf{f}_{v,\tau} \right). \quad (3)$$

To stabilize training, we employ multi-head attention, concatenating the outputs of K independent attention heads. This allows the model to capture different modes of spatial interaction (e.g., one head tracking hydrological connectivity, another tracking thermal gradients).

Temporal Dynamics (Ecological Memory). Having established the regional context $\mathbf{s}_{c,\tau}$ at each step, we next model the temporal evolution of these conditions. Ecological drivers often exhibit significant lag effects. We process the sequence of spatial features $\mathbf{S}_{c,t} = [\mathbf{s}_{c,t-W+1}, \dots, \mathbf{s}_{c,t}]$ using a Gated Recurrent Unit (GRU). At each time step τ , the update equations are:

$$\mathbf{z}_\tau = \sigma(\mathbf{W}_z \mathbf{s}_{c,\tau} + \mathbf{U}_z \mathbf{h}_{c,\tau-1} + \mathbf{b}_z), \quad (4)$$

$$\mathbf{r}_\tau = \sigma(\mathbf{W}_r \mathbf{s}_{c,\tau} + \mathbf{U}_r \mathbf{h}_{c,\tau-1} + \mathbf{b}_r), \quad (5)$$

$$\tilde{\mathbf{h}}_\tau = \tanh(\mathbf{W}_h \mathbf{s}_{c,\tau} + \mathbf{U}_h (\mathbf{r}_\tau \odot \mathbf{h}_{c,\tau-1}) + \mathbf{b}_h), \quad (6)$$

$$\mathbf{h}_{c,\tau} = (1 - \mathbf{z}_\tau) \odot \mathbf{h}_{c,\tau-1} + \mathbf{z}_\tau \odot \tilde{\mathbf{h}}_\tau. \quad (7)$$

The final hidden state $\mathbf{h}_{c,t}$ represents the *regional ecological memory*.

Point-Context Fusion. To obtain the final embedding for a specific survey location u (located within grid cell c), we fuse the high-resolution local environmental covariates $\mathbf{x}_{u,t}$ with the coarse-grained regional context $\mathbf{h}_{c,t-1}$:

$$\mathbf{h}_{u,t} = \text{MLP}(\mathbf{x}_{u,t} \parallel \mathbf{h}_{c,t-1}). \quad (8)$$

The resulting vector $\mathbf{h}_{u,t} \in \mathbb{R}^{d_h}$ serves as the final spatio-temporal habitat embedding. This hierarchical approach allows the model to distinguish between critical source habitats and irrelevant neighbors while retaining the local precision required to detect micro-habitat filters.

2.3 Context-Anchored Latent Alignment

A fundamental challenge in JSMD is the combinatorial explosion of the output space, which scales as 2^S for S species. We posit that ecologically valid assemblages are not uniformly distributed but occupy a lower-dimensional manifold

constrained by shared functional traits and environmental filters. To capture this intrinsic geometry, we employ a Conditional Variational Autoencoder (CVAE), which leverages the spatio-temporal embedding $\mathbf{h}_{u,t}$ to condition the generation of a compact latent variable $\mathbf{z}$.

Label-Activated Mixture Prior. Standard CVAEs assume a unimodal prior $p(\mathbf{z}) = \mathcal{N}(0, I)$. This is ill-suited for ecological data, which is inherently multimodal; different stable communities can exist under distinct conditions. To address this, we adopt the Label-Activated Mixture Prior [Bai *et al.*, 2022]. We assign a learnable prototype embedding $\mathbf{w}_s \in \mathbb{R}^{d_z}$ to each species s . These prototypes serve as the centroids for species-specific Gaussian components. For a given training sample with observed species set $\mathcal{P}(\mathbf{y}_{u,t}) = \{s : y_{u,t,s} = 1\}$, the prior is defined as a uniform mixture of the components associated with the *present* species:

$$p_\psi(\mathbf{z}|\mathbf{y}_{u,t}) = \frac{1}{|\mathcal{P}(\mathbf{y}_{u,t})|} \sum_{s \in \mathcal{P}(\mathbf{y}_{u,t})} \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_s, \mathbf{I} \cdot \sigma_s^2). \quad (9)$$

This formulation ensures that the prior probability mass is concentrated in regions of the latent space corresponding to valid combinations of the observed species. It creates a “gravitational field” in latent space, where the structure of the prior changes dynamically based on the observed biodiversity.

Contrastive Alignment via InfoNCE. The inference network maps the habitat embedding $\mathbf{h}_{u,t}$ to the parameters of the approximate posterior $q_\phi(\mathbf{z}|\mathbf{h}_{u,t}) = \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))$. To enable gradient descent, we employ the reparameterization trick to sample $\mathbf{z}_{u,t}$. To ensure that the sampled latent codes cluster meaningfully around their corresponding species modes, we employ a supervised contrastive objective. We first project the sampled latent code to a metric embedding space via a non-linear projection head: $\tilde{\mathbf{z}}_{u,t} = g_{\text{proj}}(\mathbf{z}_{u,t})$. We treat this projected latent sample $\tilde{\mathbf{z}}_{u,t}$ as the *anchor*. The species prototypes $\mathbf{w}_s$ act as the keys. For a given anchor, prototypes of present species are *positive keys*, while those of absent species are *negative keys*. We minimize the InfoNCE loss:

$$\mathcal{L}_{\text{con}} = - \sum_{p \in \mathcal{P}(\mathbf{y}_{u,t})} \log \frac{\exp(\text{sim}(\tilde{\mathbf{z}}_{u,t}, \mathbf{w}_p)/\tau)}{\sum_{s=1}^S \exp(\text{sim}(\tilde{\mathbf{z}}_{u,t}, \mathbf{w}_s)/\tau)}, \quad (10)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature scalar. By aligning the *stochastic* sample rather than the deterministic encoder output, we ensure that the entire volume of the posterior distribution is mapped to the correct validity region defined by the community prior.

2.4 Imbalance-Aware Decoupled Decoding

The third major component of our framework specifically addresses the “long-tail” problem inherent in biodiversity data. In standard JSMDs, the decoder typically assumes a symmetric likelihood (e.g., via Binary Cross Entropy), forcing the model to prioritize the majority class (i.e. absences) to minimize global error. Given that most species are rare, this leads to a trivial solution where the model predicts “absent” for all rare species, effectively erasing them from the predictions.

To counter this, we introduce an Imbalance-Aware Decoupled Decoding strategy. While we utilize a computationally efficient pairwise inner product as the architectural mechanism, we fundamentally alter the *probabilistic definition* of the decoder. By replacing the standard likelihood with an Asymmetric Focusing objective, we introduce a structural modification to the learning dynamics: this acts as a dynamic gradient gating mechanism that *decouples* the optimization of “easy” samples (common absences) from “hard” samples (rare presences).

Decoder Architecture (Geometric Readout). Structurally, we formulate the decoding process as a pairwise inner product between the projected latent habitat anchor $\tilde{\mathbf{z}}_{u,t}$ and the species prototypes $\mathbf{w}_s$ (the same prototypes used in the mixture prior). The probability of presence for species j is given by:

$$\hat{y}_{u,t,j} = \sigma(\tilde{\mathbf{z}}_{u,t}^\top \mathbf{w}_j), \quad (11)$$

where $\sigma(\cdot)$ is the sigmoid function. This design enforces strictly interpretable consistency: the model predicts a species is present if and only if the current spatio-temporal context embeds into that species’ specific niche region in the metric space.

Asymmetric Focusing Mechanism. To enforce sensitivity to rare species, we integrate the Asymmetric Loss (ASL) [Ridnik *et al.*, 2021] directly into the Evidence Lower Bound (ELBO). ASL operates by dynamically down-weighting the contribution of easy negatives. The reconstruction term is defined as:

$$\mathcal{L}_{\text{ASL}} = \sum_{j=1}^S -y_j L_+ - (1 - y_j) L_-, \quad (12)$$

where the positive and negative probability estimates are decoupled:

$$L_+ = (1 - p)^{\gamma_+} \log(p), \quad (13)$$

$$L_- = (p_m)^{\gamma_-} \log(1 - p_m). \quad (14)$$

Here, $p = \hat{y}_{u,t,j}$ is the predicted probability, and $p_m = \max(p - m, 0)$ is the shifted probability. The hyperparameters γ_+ and γ_- control the focusing strength. By setting $\gamma_- > \gamma_+$ (e.g., $\gamma_- = 4, \gamma_+ = 1$), we aggressively decay the loss contribution of easy negatives (clear absences) once the model is confident ($p < m$). This prevents the vast number of empty survey records from overwhelming the gradient signal derived from rare species occurrences.

Optimization. Our model is trained end-to-end by minimizing a unified objective that balances generative fidelity, latent structure alignment, and robust long-tail handling.

The total objective $\mathcal{L}_{\text{total}}$ combines the Imbalance-Aware ELBO and the auxiliary Contrastive Alignment loss.

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_q[\mathcal{L}_{\text{ASL}}(\mathbf{y}, \hat{\mathbf{y}})] - \beta \text{KL}(q_\phi(\mathbf{z}|\mathbf{h}_{u,t}) \parallel p_\psi(\mathbf{z}|\mathbf{y}_{u,t})), \quad (15)$$

where β weights the KL divergence term. The KL term forces the variational posterior (conditioned on the environment) to align with the community-derived mixture prior (conditioned on species labels). We approximate this divergence using the

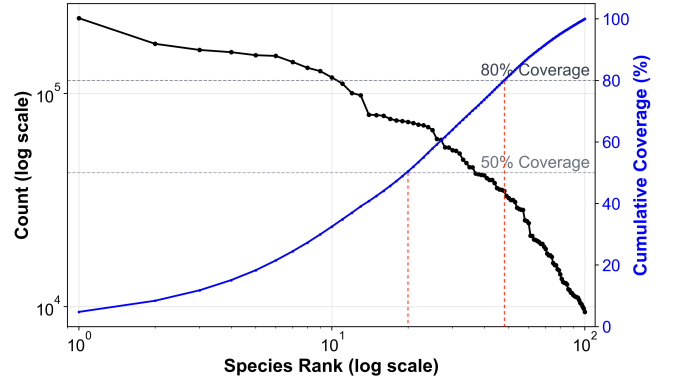


Figure 2: The Long-Tail of Biodiversity. Rank-frequency distribution of the top 100 species in the eBird dataset (log-log scale). The black line denotes species counts, while the blue curve shows cumulative coverage. The vertical dashed lines highlight extreme class imbalance: the top 20 species alone contribute 50% of the data, while the top 48 species account for 80%, leaving a long tail of rare specialists.

log-sum-exp upper bound to efficiently handle the Gaussian Mixture components defined by the species prototypes. The final loss function is:

$$\mathcal{L}_{\text{total}} = -\mathcal{L}_{\text{ELBO}} + \alpha \mathcal{L}_{\text{con}}. \quad (16)$$

3 Experiments

3.1 Dataset Curation

We evaluate our framework on the **eBird** dataset, the world’s largest citizen-science biodiversity monitoring project. eBird is a crowd-sourced bird observation dataset [Sullivan *et al.*, 2009]. We curate a subset of the eBird Basic Dataset (EBD) covering North America, restricting our analysis to the top $S = 100$ most frequently observed species. The final dataset comprises 434,352 checklists. Each checklist is associated with a 104-dimensional feature vector describing the sampling context. Consistent with ecological monitoring, the data exhibits a severe power-law distribution (see Figure 2). The top 20 dominant generalists account for 50% of all detections. Conversely, the “tail” comprises 80 specialist species, with the rarest appearing in roughly 2% of checklists.

3.2 Experimental Settings

Data Splitting. To rigorously evaluate the modeling of regional connectivity and historical dynamics, we employ a *temporal splitting strategy*. We utilize data from 2014–2018 (319,078 checklists) as *historical context* to construct the regional spatio-temporal priors. The subsequent data from 2019 (115,274 checklists) serves as the target set for model development and is randomly partitioned into training (80%), validation (10%), and testing (10%) subsets.

Baselines. We benchmark STELLAR against five state-of-the-art methods representing diverse JSDM categories. In the embedding-based domain, **DMSE** [Chen *et al.*, 2017] maps environmental features and species into a shared latent space,

Method	Example-F1	Micro-F1	Macro-F1	Precision@1	Macro Recall	Macro PR-AUC
DMSE	0.3302	0.4423	0.2707	0.4050	0.2575	0.1860
MulSupCon	0.2898	0.4148	0.2348	0.6594	0.1712	0.4207
C-GMVAE	0.3993	0.5210	0.3979	0.6943	0.3123	0.4901
SINR	0.3770	0.4060	0.2649	0.5671	0.4077	0.2233
LabelKAN	0.4148	0.5322	0.4208	0.6932	0.3420	0.4890
STELLAR	0.4895	0.5701	0.4766	0.7019	0.4911	0.4816

Table 1: Performance comparison on the eBird dataset. STELLAR consistently outperforms baselines in F1 scores, Precision@1, and Recall. While C-GMVAE achieves slightly higher PR-AUC through a conservative prediction strategy, STELLAR demonstrates superior capability in detecting rare species (Macro Recall), leading to the best overall balance between precision and coverage. *Bold indicates best performance.*

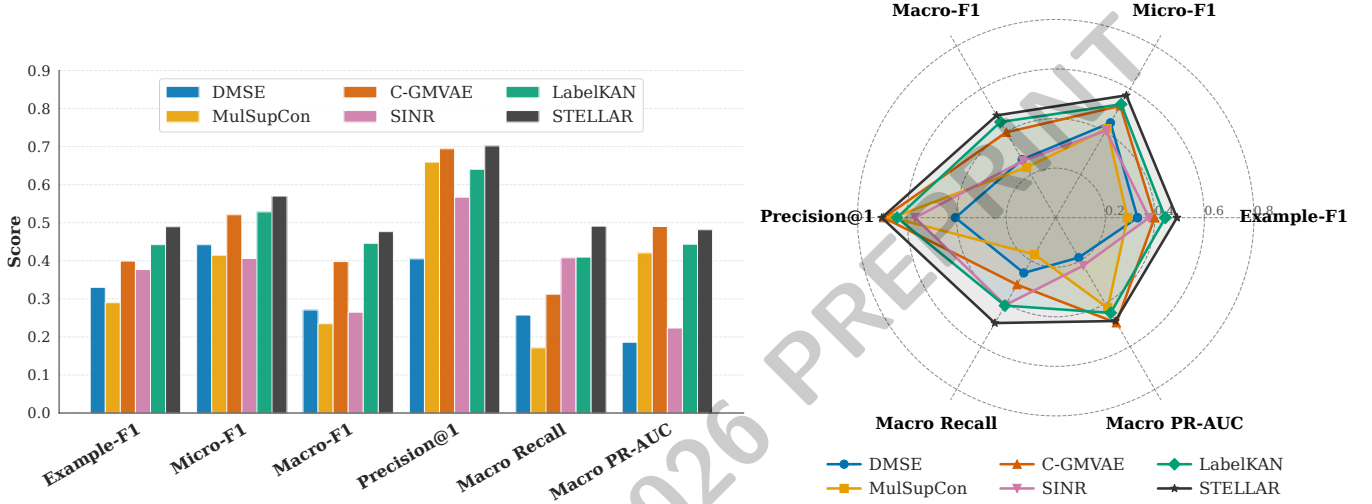


Figure 3: Visual comparison of model capabilities. The plots illustrate STELLAR’s superior handling of long-tailed distributions. By effectively balancing Precision and Recall, our framework mitigates the mode collapse observed in baselines, ensuring robust detection of both common and rare species.

modeling correlations via a multivariate probit framework. **MulSupCon** [Zhang and Wu, 2024] extends discriminative representation learning to multi-label tasks by weighting contrastive positive samples based on label overlap. Representing generative approaches, **C-GMVAE** [Bai *et al.*, 2022] conditions a VAE on environmental features using a Gaussian Mixture prior to capture multimodal distributions. For spatial representation, **SINR** [Cole *et al.*, 2023] utilizes implicit neural representations to learn continuous coordinate-based functions for global interpolation. Finally, **LabelKAN** [Grimson *et al.*, 2026] serves as a structure learning baseline, employing Kolmogorov-Arnold Networks to explicitly model high-order label correlations through learnable activation functions.

Metrics. We evaluate model performance using comprehensive multi-label classification metrics. We report **Micro-F1**, **Macro-F1**, and **Example-F1** to assess the balance between precision and recall at different aggregation levels. Crucially, **Macro-F1** and **Macro Recall** serve as the primary indicators for our long-tailed setting; by averaging performance per species regardless of support, they explicitly pe-

nalize models that fail to detect rare species. Additionally, we employ **Macro PR-AUC** and **Precision@1** to evaluate the quality of the predictive ranking and the accuracy of the top-confidence predictions, respectively.

Implementation. We implemented our model STELLAR using PyTorch and the PyG (PyTorch Geometric) library [Fey *et al.*, 2025]. All experiments were conducted on a workstation equipped with 4 NVIDIA GeForce RTX 3090 GPUs. Due to space constraints, we defer detailed specifications regarding data preprocessing, model architecture, and training strategy to the Appendix.¹

3.3 Results and Analysis

Table 1 and Figure 3 present the comparative performance on the eBird dataset. **STELLAR** achieves state-of-the-art results across the majority of metrics, surpassing the strongest baseline (LabelKAN) by a significant margin in **Macro-F1 (0.477 vs. 0.421)** and **Macro Recall (0.491 vs. 0.342)**. Notably, our model attains the highest **Precision@1 (0.702)**, demonstrating that our imbalance-aware decoding mechanism ef-

¹This paper’s appendix is available in the extended arXiv version.



Figure 4: Per-species F1 performance heatmap on the 20 rarest species (sorted by total support). The x-axis represents bird species ordered by increasing rarity (lowest support on the left). Darker cells indicate higher F1 scores. While baselines like DMSE and MulSupCon exhibit severe mode collapse (predicting zero for most rare species), STELLAR maintains robust detection capabilities across the long tail, successfully recovering species that are effectively invisible to other models.

fectively filters false positives even while maximizing recall. We observe that **C-GMVAE** yields a marginally higher **Macro PR-AUC (0.490)** compared to STELLAR (0.482). This slight difference arises because PR-AUC favors conservative models that assign high confidence to a narrow set of “easy” head classes. C-GMVAE achieves this ranking stability by sacrificing coverage, as evidenced by its significantly lower **Macro Recall (0.312)**. In contrast, STELLAR extends its predictive capacity to the difficult “long tail” of rare species. While ranking these uncertain rare events causes a minor drop in PR-AUC, it results in a massive **+57% relative improvement in Macro Recall** over C-GMVAE, proving that STELLAR offers a far more comprehensive solution for biodiversity monitoring.

Rare Species Detection Analysis. To further investigate the model’s capability in the extreme long-tail regime, we visualize the per-species F1 scores for the 20 rarest species in Figure 4. As illustrated, traditional embedding-based methods such as **DMSE** and **MulSupCon** suffer from severe mode collapse, failing to detect the majority of these species (indicated by the prevalence of 0.00 scores). This confirms that without explicit handling of class imbalance, discriminative models tend to bias heavily toward majority classes to minimize aggregate loss. While the generative baseline **C-GMVAE** improves detection rates, it still exhibits inconsistency, missing several species entirely. In contrast, **STELLAR** demonstrates superior robustness, providing consistent non-zero F1 scores across the entire set of rare species. This performance gain is directly attributable to our *Imbalance-Aware Decoupled Decoding* strategy, which prevents the gradients from dominant head classes from suppressing the learning of sparse tail signals. These results highlight STELLAR’s potential for conservation applications, where detect-

ing rare and threatened species is often the primary objective.

We also give qualitative visualizations of the spatial predictions for the 20 rarest species in the Appendix. These distribution maps reveal that STELLAR captures rare events more effectively than baselines and more accurately reconstructs fine-grained spatial structures.

Ablation Study. We conduct ablation study to quantify the contribution of each core component in STELLAR by analyzing the relative performance decline upon its removal. First, removing the Imbalance-Aware Decoupled Decoding results in a sharp 26.2% decrease in Macro Recall and a 10.3% drop in Macro-F1, verifying its critical role in recovering rare species. Next, replacing the Spatio-Temporal Encoder with a standard MLP causes a severe degradation, with a 30.4% drop in Macro Recall and an 11.7% decline in Macro-F1, confirming that explicit spatial modeling is important. Finally, simplifying the Context-Anchored Latent Alignment to a univariate Gaussian prior leads to a 5.6% decline in Macro Recall and a 3.1% drop in Macro-F1, highlighting the necessity of multimodal priors for complex distributions.

4 Conclusion

In this paper, we proposed STELLAR, a novel framework designed to address the challenges of large-scale biodiversity monitoring. Our approach integrates three core components: a GNN-RNN Encoder for capturing context-rich biophysical embeddings; a Context-Anchored Latent Alignment mechanism for multimodal structure learning; and an Imbalance-Aware Decoupled Decoding module for robust long-tail prediction. Extensive experiments and analysis confirm the superior performance of our model against state-of-the-art baselines, establishing STELLAR as a comprehensive and effective solution for fine-grained species distribution modeling.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62506090); the National Key R&D Program of China (Grant No. 2025YFF0523900); an AI2050 Senior Fellowship, a Schmidt Sciences program; the National Science Foundation (NSF); the National Institute of Food and Agriculture (USDA/NIFA; 2023-67021-39829); and the Air Force Office of Scientific Research (AFOSR; FA9550-23-1-0322).

References

- [Bai *et al.*, 2020] Junwen Bai, Shufeng Kong, and Carla Gomes. Disentangled variational autoencoder based multi-label classification with covariance-aware multivariate probit model. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.
- [Bai *et al.*, 2022] Junwen Bai, Shufeng Kong, and Carla P. Gomes. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, 2022.
- [Chen *et al.*, 2017] Di Chen, Yexiang Xue, Daniel Fink, Shuo Chen, and Carla P. Gomes. Deep multi-species embedding. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3639–3646, 2017.
- [Chen *et al.*, 2018] Di Chen, Yexiang Xue, and Carla Gomes. End-to-end learning for the deep multivariate probit model. In *International Conference on Machine Learning (ICML)*, pages 932–941. PMLR, 2018.
- [Cole *et al.*, 2023] Elijah Cole, Grant Van Horn, Christian Lange, et al. Spatial implicit neural representations for global-scale species mapping. In *International Conference on Machine Learning (ICML)*, pages 6078–6095. PMLR, 2023.
- [Elith and Leathwick, 2009] Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 40:677–697, 2009.
- [Fan *et al.*, 2022] Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P Gomes. A gnn-rnn approach for harnessing geospatial and temporal information: Application to crop yield prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11873–11881, 2022.
- [Fey *et al.*, 2025] Matthias Fey, Jinu Sunil, Akihiro Nitta, Rishi Puri, Manan Shah, Blaž Stojanovič, Ramona Bendias, Alexandria Barghi, Vid Kocijan, Zecheng Zhang, Xinwei He, Jan Eric Lenssen, and Jure Leskovec. Pyg 2.0: Scalable learning on real world graphs, 2025.
- [Golding and Purse, 2016] Nick Golding and Bethan V Purse. Fast and flexible bayesian species distribution modelling using gaussian processes. *Methods in Ecology and Evolution*, 7(5):598–608, 2016.
- [Gomes *et al.*, 2019] Carla Gomes, Thomas Dietterich, et al. Computational sustainability: Computing for a better world and a sustainable future. *Communications of the ACM*, 62(9):56–65, 2019.
- [Grimson *et al.*, 2026] Marc Grimson, Joshua Fan, Courtney Davis, Dylan van Bramer, Daniel Fink, and Carla Gomes. Labelkan - kolmogorov-arnold networks for inter-label learning: Avian community learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.
- [Lin *et al.*, 2017] 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.
- [Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.
- [Ovaskainen and Soininen, 2011] Otso Ovaskainen and Janne Soininen. Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, 92(2):289–295, 2011.
- [Ovaskainen *et al.*, 2017] Otso Ovaskainen, Gleb Tikhonov, et al. How to make more out of sparse data: hierarchical modeling of species communities. *Ecology*, 98(10):2626–2643, 2017.
- [Ridnik *et al.*, 2021] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021.
- [Sullivan *et al.*, 2009] Brian L Sullivan, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation*, 142(10):2282–2292, 2009.
- [Veličković *et al.*, 2017] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [Wisz *et al.*, 2013] Mary S Wisz, Julien Pottier, W Daniel Kissling, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological reviews*, 88(1):15–30, 2013.
- [Zhang and Wu, 2024] Pingyue Zhang and Mengyue Wu. Multi-label supervised contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16786–16793, 2024.