

FlywayBench: A Standardized Macroecological Dataset Validated via Heterogeneous Graph Neural Networks

Xinyi Chang*

School of Data Science and Intelligent Media
Communication University of China
Beijing, China
xian@cuc.edu.cn

Zexun Jiang

School of Data Science and Intelligent Media
Communication University of China
Beijing, China
jiangzx2023@cuc.edu.cn

Hongjiang Xiao

State Key Laboratory of Media Convergence and Communication
Communication University of China
Beijing, China
xiaohj@cuc.edu.cn

Abstract

Large-scale species distribution modeling (SDM) is fundamental to biodiversity conservation, yet it is severely constrained by spatial sampling bias and pseudoreplication. To bridge this gap, we introduce FlywayBench, a macroecological benchmark specifically designed for long-range spatial inference. Distilled from 30M+ raw records via an automated debiasing pipeline, FlywayBench provides 14.2 million high-quality observations for 155 species across 26 years, mapped to 4,554 grid nodes with aligned environmental covariates and Americas Flyway topological priors. To validate its utility, we propose FlywayGNN, a heterogeneous graph neural baseline that couples local adjacency with migratory connectivity. Blind extrapolation on a South American holdout shows that FlywayGNN outperforms the non-spatial baseline and achieves modest gains over Vanilla GCN. This work establishes a critical data foundation and evaluation paradigm for macroecological modeling under sparse and biased supervision. Our source code is available at <https://github.com/CHANG-1125/FlywayBench>.

CCS Concepts

• **Information systems** → **Spatial-temporal systems**; *Data mining*; • **Computing methodologies** → **Neural networks**; • **Applied computing** → **Earth and atmospheric sciences**.

Keywords

Graph Neural Networks, Species Distribution Modeling, Habitat Reconstruction, Spatial Extrapolation, Americas Flyway, Crowdsourced Data Debiasing, Earth and Atmospheric Sciences

ACM Reference Format:

Xinyi Chang, Zexun Jiang, and Hongjiang Xiao. 2026. FlywayBench: A Standardized Macroecological Dataset Validated via Heterogeneous Graph Neural Networks. In *Proceedings of The 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26) (KDD '26)*. ACM, New York, NY, USA, 8 pages.

1 Introduction

Understanding species' geographic distributions at the global scale is critical for biodiversity conservation, climate-change response, and macroecological research [16]. In recent years, crowdsourced community-science platforms such as eBird have accumulated tens

of millions of observations, providing a rich empirical foundation for Species Distribution Models (SDMs) [29]. Mining ecological signals from such globally scaled archives—and predicting species distributions and habitats, including cross-continental SDM—has become a central frontier at the intersection of data mining, machine learning, and ecology.

Although community-science platforms provide unprecedented volumes of ecological observations, robust macro-scale learning is constrained by data-quality issues. Crowdsourced archives exhibit unstructured noise, spatial heterogeneity, and sampling bias [1, 31]; disciplined data-curation pipelines—spatial filtering, de-duplication, and debiasing—are therefore needed to distill coherent spatio-temporal signals [22]. Yet when structured ecological priors are weak or absent, cross-continental SDM still faces two intertwined challenges:

- Crowdsourced observations carry geographic and socioeconomic sampling biases. Presence-only reporting creates uncertain absences, while repeated reports can introduce pseudoreplication. In sparsely observed grid cells, models easily equate *unobserved* with *true absence* and may learn the observation process rather than niche mechanisms [3].
- Classical SDMs—such as tree-based ensembles [7] and simple feedforward networks—often treat locations as i.i.d. samples and fall into the observation-effort trap, conflating human accessibility with habitat preference, which makes rigorous evaluation of spatial models difficult at continental scales [25].

To address these challenges, we construct a continental-scale macroecological benchmark dataset designed for long-range spatial extrapolation. Extracting this high-quality corpus from unstructured crowdsourced archives relies on our Automated Spatiotemporal Data Curation and Debiasing Pipeline. By applying flyway-aware vector polygon filtering and effort-aware de-duplication, the pipeline effectively mitigates spatial noise and pseudoreplication biases. This curation distills over 30 million raw eBird records (2000–2025) into 14,241,680 standardized observations, mapping them to 4,554 macroscopic $1^\circ \times 1^\circ$ grid nodes paired with aligned multi-source environmental covariates for 155 focal waterbirds. Furthermore, to validate the practical utility of this dataset and its embedded macro-topological priors, we introduce FLYWAYGNN, a heterogeneous graph neural baseline. By coupling long-range migratory connectivity with local spatial adjacency, FLYWAYGNN is designed to enrich training-domain representations with migratory

*This work is supported by the Fundamental Research Funds for the Central Universities (CUC25SG003, CUC24SG004).

topology and to evaluate extrapolation into sparse South American domains.

The core contributions are threefold:

- We assemble **FlywayBench**, a 26-year (2000–2025) spatio-temporal corpus yielding 14,241,680 high-confidence records for 155 target waterbird species. Mapping these observations to $4,554 1^\circ \times 1^\circ$ grid nodes with aligned environmental covariates provides a quantitative substrate for cross-continental SDM.
- We design the **Automated Spatiotemporal Data Curation and Debiasing Pipeline** to process massive community-science archives. Integrating flyway polygon screening and effort-aware de-duplication, it mitigates inherent observation bias and converts raw unstructured checklists into a standardized gridded ecological network.
- We develop FLYWAYGNN as a topology-aware heterogeneous graph baseline that combines local spatial adjacency with migratory-corridor priors. On the South American holdout, it provides modest aggregate improvements over the spatial GCN and reconstructs ecologically plausible habitat patterns, demonstrating that FlywayBench supports controlled comparison of local and long-range graph models.

2 Related Work

2.1 Observation Bias and Data Mining

Community-science platforms (e.g., eBird and iNaturalist) have enabled large-scale species modeling in machine learning [30]. However, these data are largely presence-only and spatially biased, so observation intensity often reflects accessibility and observer behavior rather than true ecological patterns [3]. Debiasing such records to recover reliable ecological signal is therefore a key data-mining challenge [14].

2.2 Spatial Gridding and Pipelines

Reliable benchmarks must account for distribution shift between training and test domains, as highlighted by WILDS [24]. Recent spatiotemporal data-mining pipelines in environmental settings (e.g., PM2.5-GNN and kriging-style extrapolation) combine remote-sensing covariates with spatial gridding to standardize prediction targets [20, 34]. Prior work also shows that spatiotemporal filtering, background sampling, and de-duplication are important to reduce sampling bias [26].

2.3 Spatial Benchmarks for Graph Neural Networks

GNNs are increasingly used for spatiotemporal dependency modeling and extrapolation/kriging at unobserved locations [20, 36]. Yet many spatial datasets still rely on local proximity or administrative partitions [33], which are limited for long-range ecological processes. Motivated by this gap, we construct a migration-aware benchmark from large-scale crowdsourced observations and evaluate heterogeneous GNNs on it.

Table 1: Summary of dataset construction and preprocessing pipeline.

Category	Parameter	Value
Data source	Platform	eBird
	Time window	2000-01-01–2025-12-31
Taxonomy base	Species count	WPE & IUCN Red List
Species pool	Candidate species pool size	871
Raw data	Extracted records	30,733,699
Final target	Modeled species count ($ S $)	155 target waterbird species
Final dataset	Filtered records	14,241,680

Table 2: Geographic profiles of the Americas Flyway System.

Route	Longitude	Latitude	Area
Pacific	$[-180.0^\circ \sim -63.7^\circ]; [152.2^\circ \sim 180.0^\circ]$	$-56.1^\circ \sim 73.9^\circ$	21.30
Central	$-130.3^\circ \sim -50.9^\circ$	$-55.0^\circ \sim 80.0^\circ$	30.45
Atlantic	$-107.5^\circ \sim -33.2^\circ$	$-56.0^\circ \sim 82.1^\circ$	33.89

3 Automated Spatiotemporal Data Curation and Debiasing Pipeline

This section presents the proposed Automated Spatiotemporal Data Curation and Debiasing Pipeline, built upon 26 years of unstructured observation records from eBird [9]—a global community-science platform operated by the Cornell Lab of Ornithology [29]. To bridge biased crowdsourced point clouds and machine-learning-ready inputs, we implement a four-stage workflow: (1) vector polygon spatial filtering to retain flyway-relevant records; (2) macroscopic spatial gridding to discretize observations; (3) effort-aware debiasing to mitigate the "group birding" effect and pseudoreplication; and (4) multi-scale environmental feature alignment to provide ecological context. This pipeline transforms massive, irregular archives into a standardized, graph-ready spatiotemporal substrate.

3.1 Data Acquisition and Species Definition

As summarized in Table 1, taxonomy follows the "Waterbird Population Estimates" (WPE) standard [10] and is cross-validated against the BirdLife-assessed IUCN Red List [5]. Using species scientific names within a fixed time window, we initially retrieved over 30 million raw observation records.

We use the Americas Flyway System as a macro-spatial topological prior. Specifically, we apply Americas flyway geographic polygons sourced from BirdLife International’s spatial conservation layers (BirdLife Data Zone) [4], distributed under their applicable terms, in the WGS84 coordinate reference system (EPSG:4326). To eliminate out-of-domain noise, we prescreen the initial point clouds using a global bounding box (longitude $[-180.00^\circ, 180.00^\circ]$, latitude $[-56.11^\circ, 82.08^\circ]$), followed by a rigorous point-in-polygon filtering process based on equal-area summaries. Table 2 reports the corridor ranges and projected areas, and Figure 2 illustrates the corresponding subregions.

At the species level, we cross-match flyway-covered taxa with eBird-observed taxa [9] by exact scientific name. This refinement

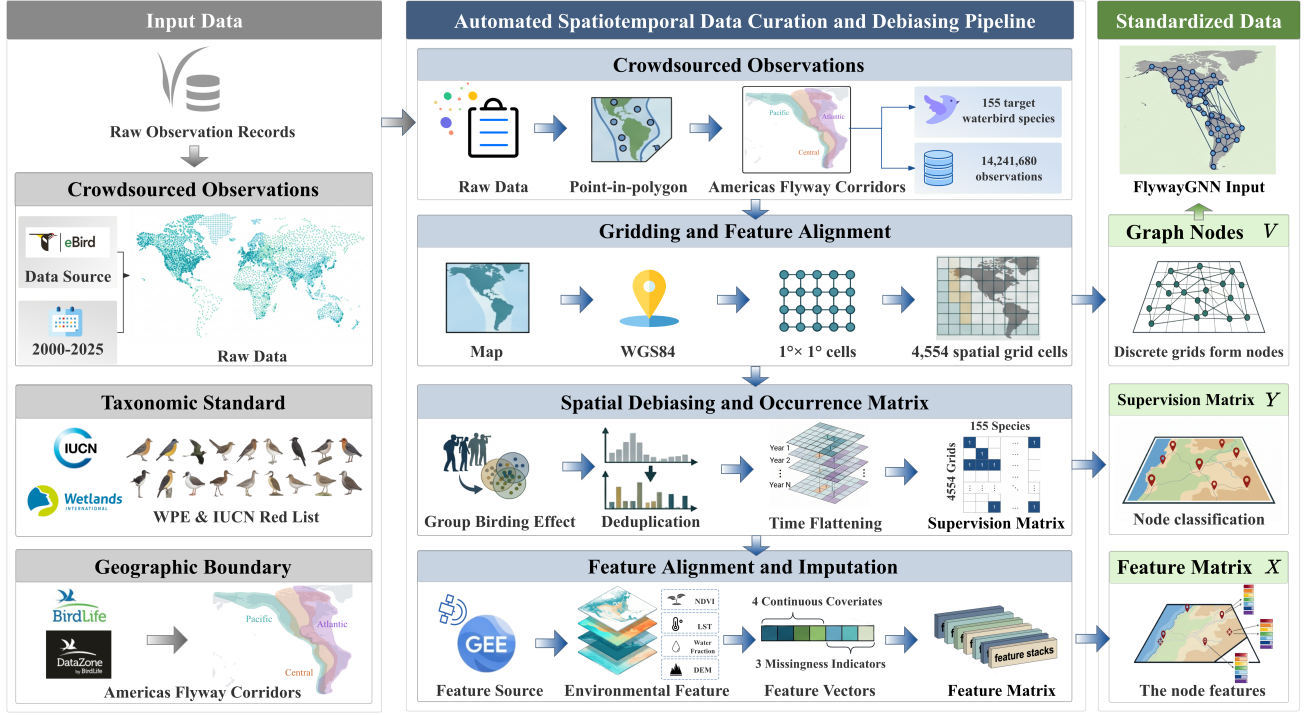


Figure 1: Overview of the automated spatiotemporal data curation and debiasing pipeline.

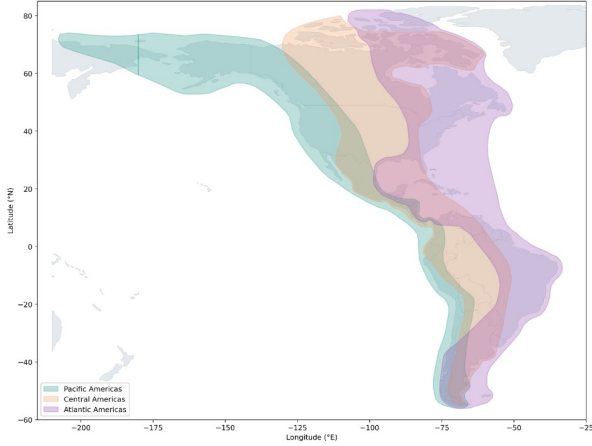


Figure 2: Americas flyway subregions for spatial filtering and topology priors.

yields 14,241,680 observations within Americas flyway corridors and a final modeling set S with $|S| = 155$.

3.2 Gridding and Feature Alignment

To convert irregular observations into graph-ready nodes, we perform standardized gridding over the Americas flyway corridors under WGS84. The region is partitioned into fixed $1^\circ \times 1^\circ$ cells (about $110 \text{ km} \times 110 \text{ km}$ near the equator, with area decreasing

toward higher latitudes) [19]. Using a floor operation, each observation is mapped to a unique grid identifier:

$$i = (c_\phi, c_\lambda) = (\lfloor \phi \rfloor, \lfloor \lambda \rfloor). \quad (1)$$

Here, ϕ and λ denote raw latitude and longitude, and c_ϕ, c_λ are grid indices. This topological discretization step substantially reduces the dimensionality of massive point-cloud data and extracts 4,554 spatial grid cells with valid historical observations across the Americas flyway region.

3.3 Spatial Debiasing and Occurrence Matrix

Although eBird’s standardized checklist mechanism provides broad coverage, it is susceptible to Group Birding Effect: repeated reports in the same spatiotemporal context can inflate abundance and effort. Following eBird-consistent principles [29], we apply spatial-temporal deduplication. After gridding, let $\mathcal{R}$ be the raw record set and each record be $r = (s, i, t, c)$; for each group with identical (s, i, t) , denoted $\mathcal{R}_{i,s,t}$, we use the maximum count as the daily abundance estimator:

$$\hat{c}_{i,s,t} = \max_{r \in \mathcal{R}_{i,s,t}} (c_r). \quad (2)$$

For blind-count entries, we retain numeric counts ($c > 0$) and map presence-only markers ($c = 'X'$) to 1 [21]. We then flatten the temporal dimension within the 26-year window T (2000–2025) and define static occurrence labels:

$$y_{i,s} = \mathbb{I} \left(\sum_{t \in T} \hat{c}_{i,s,t} > 0 \right). \quad (3)$$

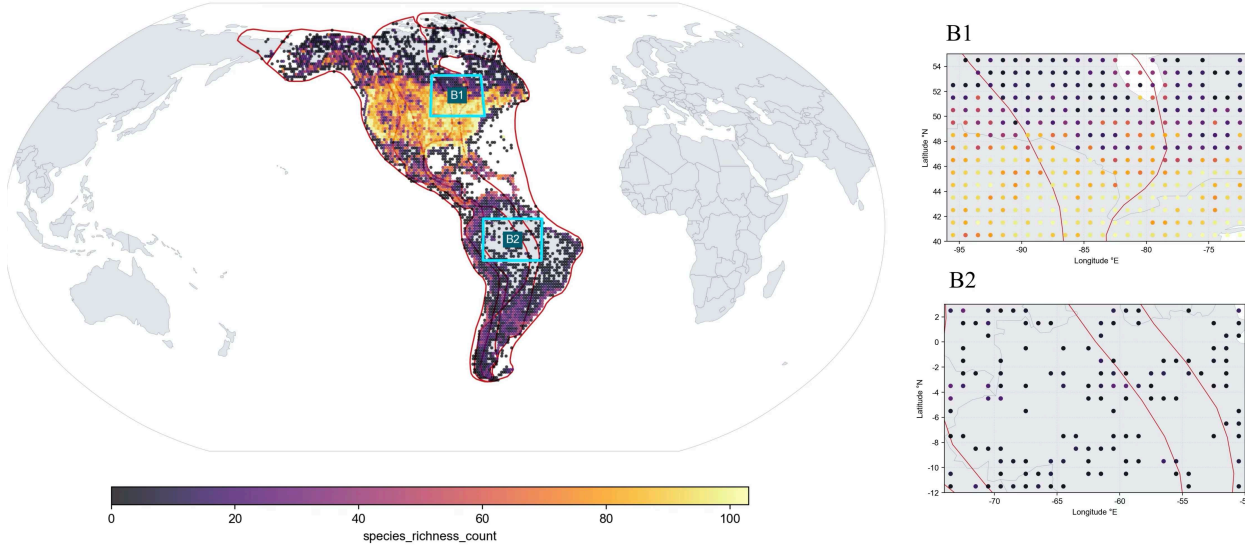


Figure 3: Gridded species-richness overview with regional zoom-ins (B1/B2) after spatial debiasing.

A species is marked present if any valid signal appears in the full time window, yielding $Y \in \{0, 1\}^{4554 \times 155}$ for graph learning.

3.4 Feature Alignment and Imputation

Using Google Earth Engine [15], we retrieve four grid-aligned indicators over 2000–2025: (1) mean NDVI from MOD13A2 [11]; (2) daytime land-surface temperature from MOD11A2 [32]; (3) mean elevation from the Copernicus GLO-30 digital elevation model [12]; and (4) water fraction from the JRC Global Surface Water maxextent layer (30 m; area-normalized to $[0, 1]$) [28].

For missing NDVI/LST values, we apply median imputation estimated during preprocessing on the assembled feature matrix (not restricted to the geographic training split) and append missingness indicators. DEM missingness (mainly open-ocean cells) is imputed with 0m using the same indicator strategy, while maxextent requires no imputation. Each node receives a feature vector $\mathbf{x}_i \in \mathbb{R}^7$ (four continuous covariates plus three missingness indicators), forming aligned $X \in \mathbb{R}^{4554 \times d}$ and supervision $Y \in \{0, 1\}^{4554 \times m}$, and supporting subsequent heterogeneous SDM graph modeling [18].

3.5 Dataset Summary

Through the proposed pipeline, we construct FlywayBench, providing a standardized foundation for cross-continental SDM. The final corpus includes:

- **Scale & Scope:** 14,241,680 high-confidence observations (distilled from 30,733,699 raw records) spanning a 26-year window (2000–2025) for 155 focal waterbird species.
- **Spatial Graph Basis:** 4,554 macroscopic $1^\circ \times 1^\circ$ grid nodes distributed across the Americas flyway corridors.
- **Node Features (X):** A 7-dimensional feature vector per node, integrating core environmental covariates (NDVI, LST, DEM, and Water Fraction) and missingness indicators.

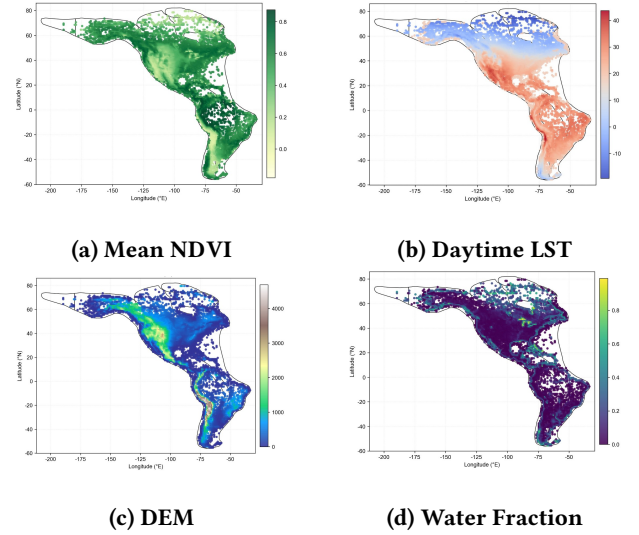


Figure 4: Multi-source environmental covariates at $1^\circ \times 1^\circ$ resolution: (a) NDVI, (b) daytime LST, (c) elevation, and (d) water fraction, used as node features in the graph.

- **Supervision Matrix (Y):** A static binary occurrence matrix $Y \in \{0, 1\}^{4554 \times 155}$, capturing effort-debiased species presence across the full temporal window.

Taken together, this curated corpus reduces unstructured spatial bias and pseudoreplication relative to raw checklists, yielding a graph-ready substrate for evaluating long-range spatial inference.

4 FLYWAYGNN: A Heterogeneous Graph Neural Framework with Migratory Topological Priors

To explicitly validate the utility of the proposed dataset and its embedded macroecological priors, we introduce FLYWAYGNN, a heterogeneous graph neural framework. To tackle cross-continental extrapolation under sparse supervision, FLYWAYGNN employs a dual-channel topology that couples local spatial adjacency with long-range migratory connectivity. Through heterogeneous message passing, it learns training-domain representations using migratory topology and evaluates them under South American extrapolation.

4.1 Graph Construction

4.1.1 Graph Formulation. Let $\mathcal{V} = \{v_i\}_{i=1}^N$ denote the valid grid-node set, where $N=4554$. We define the graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X}). \quad (4)$$

Here, $\mathcal{E}$ denotes heterogeneous edges for local spatial proximity and macro-scale flyway connectivity (Section 4.2). The matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ stacks node-feature rows $\mathbf{x}_i \in \mathbb{R}^d$ (climatic, vegetation, hydrological, and terrain covariates); Section 3.4 sets $d = 7$.

4.1.2 Debiasing-Aware Supervision. We formulate intercontinental species-distribution inference as multi-label node classification on $\mathcal{G}$. Let $\mathcal{S}$ be the target species set with cardinality $|\mathcal{S}| = m = 155$. We define the observation label matrix $\mathbf{Y} \in \{0, 1\}^{N \times m}$, where $y_{i,s} = 1$ indicates species s is observed in node v_i , and 0 otherwise. To handle uncertain and missing labels in crowdsourced data, we introduce a supervision mask matrix $\mathbf{M} \in \{0, 1\}^{N \times m}$ following masked-loss principles [8].

- **High-confidence entries** ($m_{i,s} = 1$): the node-species pair (v_i, s) has reliable positive evidence, or is treated as a reliable negative after effort-aware debiasing; gradients are enabled.
- **Uncertain missing entries** ($m_{i,s} = 0$): the pair lacks sufficient effort support and remains status-unknown. The term is masked out so that unobserved entries are not treated as true absences.

4.2 Hybrid Topology Construction

To jointly encode local spatial autocorrelation and long-range migratory connectivity, we define the edge set as

$$\mathcal{E} = \mathcal{E}_S \cup \mathcal{E}_F. \quad (5)$$

4.2.1 Spatial Adjacency Edges ($\mathcal{E}_S$). Spatial edges are built from the Moore 8-neighborhood on the 1° grid: if v_i and v_j are side-adjacent or diagonally adjacent, then $(i, j) \in \mathcal{E}_S$. These edges encode local spatial autocorrelation in climate, vegetation cover, and habitat continuity.

4.2.2 Flyway Connectivity Edges ($\mathcal{E}_F$). Flyway edges encode long-range migratory priors between breeding- and wintering-side hubs within shared corridors. We first map each node via point-in-polygon operations to BirdLife flyway corridors [4], with corridor(v_i) $\in$ {Pacific, Central, Atlantic}. For each species s , a pair (v_i, v_j) enters the species-specific set $\mathcal{E}_F^{(s)}$ if (i) $v_i, v_j \in \mathcal{V}_{\text{train}}$ and $y_{i,s} = y_{j,s} = 1$, (ii) breeding/wintering latitude constraints hold (e.g., $\phi_i \geq \phi_B$ and

$\phi_j \leq \phi_W$), and (iii) the nodes share a corridor. To avoid over-dense connectivity [23], we retain only Top- K hubs per side within each species-corridor group, scored by 26-year observation frequency on $\mathcal{V}_{\text{train}}$, following hub-node and site-fidelity motivations [2, 13, 27]. The shared flyway graph used by the dual-channel encoder is

$$\mathcal{E}_F = \bigcup_{s \in \mathcal{S}} \mathcal{E}_F^{(s)}. \quad (6)$$

Thus $\mathcal{E}_F$ links training-domain hubs and does not attach flyway edges to $\mathcal{V}_{\text{test}}$. Test-domain nodes remain unlabeled structural nodes and receive flyway-informed representations mainly through subsequent spatial mixing on $\mathcal{E}_S$.

4.3 Heterogeneous Message Passing and Feature Fusion

Because $\mathcal{E}_S$ and $\mathcal{E}_F$ represent distinct ecological semantics, FLYWAYGNN uses dual-channel heterogeneous message passing. We first map $\mathbf{X}$ to initial hidden states via an encoder MLP: $\mathbf{H}^{(0)} = \text{MLP}_{\text{enc}}(\mathbf{X})$. At layer l , we compute channel-wise aggregated messages:

$$\begin{aligned} \mathbf{M}_S^{(l)} &= \text{AGG}_S(\mathbf{H}^{(l-1)}, \mathcal{E}_S), \\ \mathbf{M}_F^{(l)} &= \text{AGG}_F(\mathbf{H}^{(l-1)}, \mathcal{E}_F). \end{aligned} \quad (7)$$

Both aggregators adopt the symmetric normalized GCN operator of Kipf and Welling [23]. For the spatial channel:

$$\mathbf{m}_{S,i}^{(l)} = \sum_{j \in \mathcal{N}_S(i)} \frac{1}{\sqrt{d_{S,i}d_{S,j}}} \mathbf{h}_j^{(l-1)}, \quad (8)$$

where $\mathcal{N}_S(i)$ is the Moore neighborhood excluding i (neighbor-only convolution), $d_{S,i} = |\mathcal{N}_S(i)|$, and the node's own embedding is carried by the explicit self-map $\mathbf{H}^{(l-1)} \mathbf{W}_{\text{self}}^{(l)}$ in Eq. (10) (analogously for the flyway channel). Inspired by semantic-level fusion in heterogeneous attention networks [35], we introduce learnable channel weights:

$$[\alpha_S^{(l)}, \alpha_F^{(l)}] = \text{softmax}([a_S^{(l)}, a_F^{(l)}]), \quad \alpha_S^{(l)} + \alpha_F^{(l)} = 1. \quad (9)$$

Node states are updated via channel-specific affine transforms plus a residual branch to alleviate over-smoothing [6]:

$$\mathbf{H}^{(l)} = \sigma\left(\alpha_S^{(l)} \mathbf{M}_S^{(l)} \mathbf{W}_S^{(l)} + \alpha_F^{(l)} \mathbf{M}_F^{(l)} \mathbf{W}_F^{(l)} + \mathbf{H}^{(l-1)} \mathbf{W}_{\text{self}}^{(l)}\right). \quad (10)$$

4.4 Masked Training for Debiasing

The output layer applies sigmoid activation to produce multi-label habitat probabilities $\hat{\mathbf{Y}} = \sigma(\mathbf{Z}) \in (0, 1)^{N \times m}$. On top of $\mathbf{M}$, we define a domain-supervision mask $d_{v,s} \in \{0, 1\}$, where $d_{v,s} = 1$ iff $v \in \mathcal{V}_{\text{train}}$ and $m_{v,s} = 1$. Let $Z := \sum_{v \in \mathcal{V}} \sum_{s \in \mathcal{S}} d_{v,s}$ denote the total mass of active domain-supervision masks. The final objective is masked binary cross-entropy averaged over valid supervision entries:

$$\mathcal{L} = -\frac{1}{Z} \sum_{v \in \mathcal{V}} \sum_{s \in \mathcal{S}} d_{v,s} \left[y_{v,s} \log \hat{y}_{v,s} + (1 - y_{v,s}) \log(1 - \hat{y}_{v,s}) \right]. \quad (11)$$

This objective restricts supervision to the training domain and prevents target-domain labels from contributing to the loss, consistent with paradigms in missing-label learning and inductive graph learning [8, 17]. At the same time, test-domain nodes can still participate in forward propagation as unlabeled structural context, while

Z-score normalization statistics are estimated from the training domain.

5 Evaluation

To comprehensively validate the foundational utility of the curated dataset and the practical application value of FLYWAYGNN in a large-scale real ecological system, we design multi-dimensional qualitative and quantitative experiments.

5.1 Experimental Setup

We study multi-label node classification on the Americas flyway graph with $N = 4554$ nodes and $m = 155$ target species. To test spatial extrapolation explicitly, we use a geographical split with latitude threshold $\tau = -12^\circ$ and center latitude $\phi(v)$ for each 1° grid node v :

- $\mathcal{V}_{\text{train}}$ ($N = 3733$): nodes with $\phi(v) \geq -12^\circ$, covering North America, Central America, and northern South America with relatively dense observations.
- $\mathcal{V}_{\text{test}}$ ($N = 821$): nodes with $\phi(v) < -12^\circ$, corresponding to sparse interior/southern South America and used as a pure extrapolation target.

For ablation and baseline comparison, we use:

- **XGBoost**: a non-spatial IID baseline using only local remote-sensing node features.
- **Vanilla GCN**: a spatial homogeneous-graph baseline that keeps only local adjacency edges $\mathcal{E}_S$ and removes flyway edges $\mathcal{E}_F$.

All Z-score normalization statistics are estimated strictly from $\mathcal{V}_{\text{train}}$ to prevent feature leakage from the test domain, whereas median imputation is fixed in preprocessing on the assembled feature matrix. Within $\mathcal{V}_{\text{train}}$, we hold out a random validation subset of nodes for hyperparameter selection; $\mathcal{V}_{\text{test}}$ is never used for model selection. Experiments are implemented in PyTorch, hyperparameters are chosen by grid search on validation macro-AUPRC, and each setting is repeated with 30 random seeds; we report mean $\pm$ standard deviation.

5.2 Habitat Reconstruction

Because sparse labels can obscure metric-only interpretation, we also assess ecological plausibility in the South American extrapolation domain ($N = 821$) by aggregating species-wise predictions into species-richness maps. XGBoost and the GNN baselines output probabilities on different calibration scales; we apply the same per-node richness aggregation to those probabilities, so maps are comparable in spatial pattern even though absolute richness means can differ sharply (e.g., 2.39 vs. 47). The non-spatial XGBoost baseline produces fragmented predictions (mean richness 2.39; 99.5th-percentile hotspot 5.02). Vanilla GCN raises mean richness to 42.79 but exhibits isotropic over-diffusion. With flyway constraints, FLYWAYGNN produces mean richness of 47.82 and raises the hotspot value to 90.27 (vs. 87.84 for Vanilla GCN), while reconstructing a continuous high-probability corridor in the La Plata Basin. These patterns are consistent with smoother cross-hemisphere habitat reconstruction, although separating ecological signal from residual observation bias remains challenging.

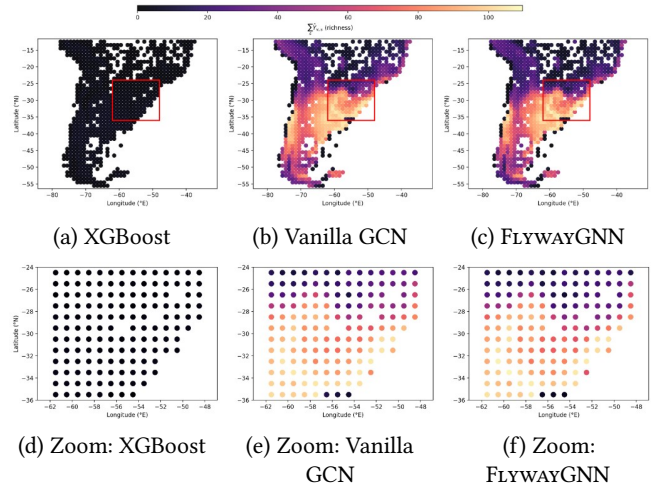


Figure 5: Richness reconstruction in the South American extrapolation domain (global and La Plata zoom-ins).

Table 3: Quantitative performance comparison on the South American extrapolation domain (eval-m5 subset, $|S_{\text{eval}}| = 73$), mean $\pm$ std over 30 seeds.

Model	Macro-AUPRC	Macro-AUC	Win/Tie/Loss
FLYWAYGNN	0.319 $\pm$ 0.006	0.695 $\pm$ 0.007	47/0/26
Vanilla GCN	0.317 $\pm$ 0.008	0.689 $\pm$ 0.012	Baseline
XGBoost	0.293 $\pm$ 0.003	0.667 $\pm$ 0.003	–

5.3 Quantitative Evaluation

We compare models in the extrapolation domain. Macro scores over all 155 species are diluted by labels with almost no test positives; we report **eval-m5**, species with ≥ 5 positives in the test domain ($|S_{\text{eval}}| = 73$).

Table 3: FLYWAYGNN obtains the highest eval-m5 scores, although its aggregate advantage over Vanilla GCN is modest (Macro-AUPRC: 0.319 vs. 0.317). It improves species-level AUPRC for 47 of 73 evaluable species, suggesting heterogeneous benefits across species.

6 Conclusion

We present FlywayBench, a long-range spatial benchmark distilled from crowdsourced observations through an automated curation and debiasing pipeline. Its 14,241,680 records form a structured environmental-graph matrix for 155 waterbird species. As a topology-aware validation baseline, FLYWAYGNN outperforms the non-spatial model and obtains modest gains over Vanilla GCN on the South American holdout, improving 47 of 73 evaluable species. These heterogeneous outcomes motivate future analysis of which species benefit most from migratory topology. Overall, FlywayBench provides a practical data foundation and evaluation protocol for macroecological modeling under sparse and biased supervision.

References

- [1] S. Balachandrar, S. Sadhuka, B. Berger, et al. 2026. Urban Incident Prediction with Graph Neural Networks: Integrating Government Ratings and Crowdsourced Reports. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 38194–38204.
- [2] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.
- [3] Sara Beery, Elijah Cole, Jason Parker, Vikram Srinivasan, Pietro Perona, and Kevin Winner. 2021. Species Distribution Modeling for Machine Learning Practitioners: A Review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '21)*. ACM, New York, NY, USA, 329–348.
- [4] BirdLife International. 2023. Americas flyway boundary spatial data (BirdLife Data Zone). <https://datazone.birdlife.org/>. Geographic polygons used for corridor filtering; use subject to BirdLife data terms.
- [5] BirdLife International. 2023. IUCN Red List for Birds species assessments (BirdLife Data Zone). <https://datazone.birdlife.org/species/search>.
- [6] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaile Li. 2020. Simple and Deep Graph Convolutional Networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*. PMLR, 1725–1735.
- [7] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794.
- [8] Elijah Cole, Oisín Mac Aodha, Titouan Lorieul, Pietro Perona, Max Welling, and Serge Belongie. 2021. Multi-Label Learning from Single Positive Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 933–942.
- [9] Cornell Lab of Ornithology. 2026. eBird: Basic Dataset [database]. <https://ebird.org/home>. Community-science observations retrieved under eBird data terms for research use.
- [10] Simon Delany, Derek A. Scott, and Adriaan T. F. Helmink. 2006. *Waterbird Population Estimates*. Wetlands International, Wageningen, The Netherlands.
- [11] Kamel Didan. 2021. MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V061. NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC). Dataset MOD13A2.061, MODIS Collection 6.1.
- [12] European Space Agency. 2021. Copernicus Digital Elevation Model (GLO-30). <https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model>. Copernicus Programme; accessed via Google Earth Engine COPERNICUS/DEM/GLO30.
- [13] Daniel Fink, Tom Auer, Alison Johnston, Matt E. Strimas-Mackey, Shawn Ligocki, Orin J. Robinson, Wesley M. Hochachka, Lisa Jaromczyk, Chad Crowley, Kylee Dunham, Andrew Stillman, Ian Davies, Amanda D. Rodewald, Viviana Ruiz-Gutierrez, and Chris Wood. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications* 30, 3 (2020), e02056.
- [14] Elise M. Geurts, John D. Reynolds, and Brian M. Starzomski. 2023. Turning observations into biodiversity data: Broad-scale spatial biases in community science. *Ecosphere* 14, 6 (2023), e4582.
- [15] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment* 202 (2017), 18–27. doi:10.1016/j.rse.2017.06.031
- [16] Antoine Guisan and Carsten Rahbek. 2011. SESAM—a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography* 38, 8 (2011), 1433–1444.
- [17] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems*, Vol. 30.
- [18] Lauren Harrell, Christine Kaeser-Chen, Burcu Karagol Ayan, et al. 2025. Heterogeneous Graph Neural Networks for Species Distribution Modeling. ICLR 2025 Workshop on Tackling Climate Change with Machine Learning.
- [19] Bradford A. Hawkins, Richard Field, Howard V. Cornell, David J. Currie, Jean-Francois Guegan, Dawn M. Kaufman, Jeremy T. Kerr, Gary G. Mittelbach, Thierry Oberdorff, Edward M. O'Brien, Eric E. Porter, and John R. G. Turner. 2003. Energy, water, and broad-scale geographic patterns of species richness. *Ecology* 84, 12 (2003), 3105–3117.
- [20] Junfeng Hu, Yuxuan Liang, Zhencheng Fan, Hongyang Chen, Yu Zheng, and Roger Zimmermann. 2023. Graph Neural Processes for Spatio-Temporal Extrapolation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. ACM, 752–763.
- [21] Alison Johnston, Daniel Fink, Monica D. Reynolds, Wesley M. Hochachka, Brian L. Sullivan, Nicholas E. Bruns, Alphonso Martínez-Abraín, David C. Pavlacky Jr., Kenneth V. Rosenberg, and Steve Kelling. 2015. Abundance models improve spatial and temporal prioritization of conservation resources. *Ecological Applications* 25, 7 (2015), 1749–1756.
- [22] Alison Johnston, Wesley M. Hochachka, Matt E. Strimas-Mackey, Viviana Ruiz-Gutierrez, Orin J. Robinson, Eliot T. Miller, Tom Auer, Steve Kelling, and Daniel Fink. 2021. Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions* 27, 7 (2021), 1265–1277.
- [23] Thomas N. Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907.
- [24] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, et al. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*. PMLR, 5637–5664.
- [25] Julie A. Lee-Yaw, Jenny L. McCune, Samuel Pironon, and Seema N. Sheth. 2022. Species distribution models rarely predict the biology of real populations. *Ecography* 2022, 6 (2022), e05877.
- [26] Bin Lu, Xiaoying Gan, Wei Zhang, et al. 2022. Spatio-Temporal Graph Few-Shot Learning with Cross-City Knowledge Transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. ACM, 1162–1172.
- [27] Ian Newton. 2023. *The Migration Ecology of Birds*. Elsevier.
- [28] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 7633 (2016), 418–422.
- [29] Brian L. Sullivan, Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (2009), 2282–2292.
- [30] M. Teng, A. Elmustafa, B. Akera, et al. 2023. SatBird: A Dataset for Bird Species Distribution Modeling Using Remote Sensing and Citizen Science Data. In *Advances in Neural Information Processing Systems*, Vol. 36. 75925–75950.
- [31] Yongxin Tong, Zhou Zhou, Yuxiang Zeng, et al. 2020. Spatial Crowdsourcing: A Survey. *The VLDB Journal* 29, 1 (2020), 217–250.
- [32] Z. Wan. 2021. MOD11A2 MODIS/Terra Land Surface Temperature and Emissivity Daily L3 Global 1km SIN Grid V061. NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC). Dataset MOD11A2.061 (MODIS Collection 6.1); accessed via Google Earth Engine MODIS/061/MOD11A2.
- [33] J. Wang, Q. Hao, W. Huang, et al. 2025. CoopRide: Cooperate All Grids in City-Scale Ride-Hailing Dispatching with Multi-Agent Reinforcement Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Vol. 1. ACM, 1457–1468.
- [34] S. Wang, Y. Li, J. Zhang, et al. 2020. PM2.5-GNN: A Domain Knowledge Enhanced Graph Neural Network for PM2.5 Forecasting. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. ACM, 163–166.
- [35] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Philip Cui, and Philip S. Yu. 2019. Heterogeneous Graph Attention Network. In *Proceedings of the World Wide Web Conference (WWW '19)*. ACM, 2022–2032.
- [36] Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, Jianzhong Qi, Chaochao Chen, and Longbiao Chen. 2023. INCREASE: Inductive Graph Representation Learning for Spatio-Temporal Kriging. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. ACM, 673–683.

A Appendix

A.1 Method overview figure

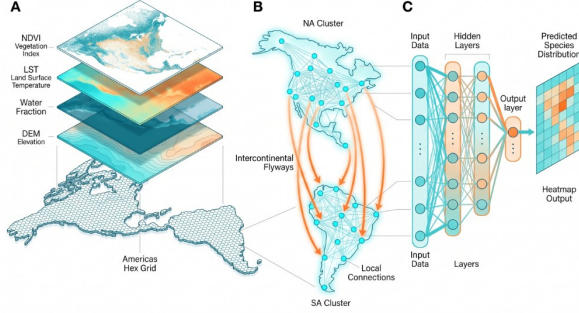


Figure 6: Overview of the proposed FLYWAYGNN framework.

A.2 Hyperparameter search space

Table 4: Hyperparameter search space and rationale.

Hyperparameter	Candidates	Role	Rationale
Learning rate (lr)	$\{10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}\}$	Optimization step size	Standard log-scale search
Hidden dimension ($hidden$)	$\{64, 128, 256\}$	Model capacity	Small/medium/large capacity
Core hub count (K)	$\{3, 5, 8, 10\}$	Flyway graph sparsity	Method-specific structural prior

A.3 Best validation configurations and parameter counts

Table 5: Best validation configurations and parameter counts (30 seeds).

Method	lr	$hidden$	K	Val. Metric (mean $\pm$ std)	# Parameters (M)
FLYWAYGNN (Ours)	0.005	256	3	0.630 $\pm$ 0.011	0.481
Vanilla GCN	0.005	256	–	0.626 $\pm$ 0.010	0.350
XGBoost	0.1	–	–	0.593 $\pm$ 0.008	–

A.4 Data split and evaluation protocol

We follow a fixed geography-based extrapolation setup. Let the node center latitude be $\phi(v)$ and threshold $\tau = -12^\circ$. We define:

$$\mathcal{V}_{\text{train}} = \{v : \phi(v) \geq \tau\}, \quad \mathcal{V}_{\text{test}} = \{v : \phi(v) < \tau\}. \quad (12)$$

Training supervision is restricted to $\mathcal{V}_{\text{train}}$, while $\mathcal{V}_{\text{test}}$ is used only for evaluation.

Because many species in the South American extrapolation domain have very few positives, main quantitative conclusions use the evaluable subset

$$\mathcal{S}_{\text{eval}} = \{s : \text{\#positive labels of } s \text{ in } \mathcal{V}_{\text{test}} \geq 5\}, \quad |\mathcal{S}_{\text{eval}}| = 73 \quad (13)$$

(eval-m5).

All results are reported over $S = 30$ random seeds. For metric t_s under seed s , we report:

$$\mu = \frac{1}{S} \sum_{s=1}^S t_s, \quad \sigma = \sqrt{\frac{1}{S-1} \sum_{s=1}^S (t_s - \mu)^2}. \quad (14)$$

Thus, all table entries are shown as $\mu \pm \sigma$.

A.5 Model selection, training schedule, and complexity

We perform grid search on validation macro-AUPRC using the search space listed in Table 4. The validation partition is a random held-out subset of nodes from $\mathcal{V}_{\text{train}}$ and is disjoint from $\mathcal{V}_{\text{test}}$:

$$lr \in \{10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}\},$$

$$hidden \in \{64, 128, 256\},$$

$$K \in \{3, 5, 8, 10\}.$$

The selected FLYWAYGNN configuration is ($lr = 0.005$, $hidden = 256$, $K = 3$). For fair comparison, Vanilla GCN uses the same lr and hidden size (with no flyway K), while XGBoost follows fixed tree hyperparameters as described in the main experiment script. Approximate parameter counts (selected setting) are reported in Table 5: FLYWAYGNN 0.481M and Vanilla GCN 0.350M. Training uses full-graph mode (one whole-graph batch per epoch, no mini-batching): max epochs are 45 (grid) and 80 (final), with early-stopping patience 8 and 12, respectively.

A.6 Extended metrics on the eval-m5 subset

Metric definitions (eval-m5):

- **Macro-AUCROC:** macro average of per-species ROC-AUC over $\mathcal{S}_{\text{eval}}$.
- **Win/Tie/Loss vs. Vanilla:** counts of species where FLYWAYGNN AUPRC is higher/equal/lower than Vanilla GCN.

Table 6: Extended eval-m5 results ($\mu \pm \sigma$, 30 seeds).

Method	Macro-AUPRC	Macro-AUCROC	Win/Tie/Loss
FLYWAYGNN	0.319 $\pm$ 0.006	0.695 $\pm$ 0.007	47/0/26
Vanilla GCN	0.317 $\pm$ 0.008	0.689 $\pm$ 0.012	–
XGBoost	0.293 $\pm$ 0.003	0.667 $\pm$ 0.003	–

Paired one-sided Wilcoxon test for macro-AUPRC on eval-m5 (FLYWAYGNN vs. Vanilla GCN): $p = 0.037$.

A.7 Reproducibility checklist and computational cost

For reproducibility, we report environment settings and computational cost:

- **Software environment:** Python 3.10+, PyTorch 2.x, PyG, scikit-learn, XGBoost, NumPy, Pandas, GeoPandas/Cartopy.
- **Randomness control:** a fixed seed list of 30 integers is used consistently across FLYWAYGNN, Vanilla GCN, and XGBoost.
- **Selection protocol:** hyperparameters are chosen strictly on validation data; test-domain metrics are reported only after model selection.
- **Compute note:** results in this version are produced under CPU execution logs; GPU configuration can be used for acceleration but does not change the evaluation protocol.
- **Reporting:** we provide both per-seed records and aggregated $\mu \pm \sigma$, plus paired significance tests.