A Niche in the Machine: The Promise of AI Foundation Models for Species Distribution Modeling

Authors: Russell Dinnage[1,2], Dan L. Warren[3]

Affiliations:
1) Department of Biological Sciences, University of Alberta, Edmonton, Canada
2) Alberta Machine Intelligence Institute, Edmonton, Canada
3) Gulbali Institute, Charles Sturt University, Thurgoona, Australia

Abstract

Species distribution models (SDM) are fundamental tools for conservation, yet methodological progress has stalled. Despite two decades of refinement, traditional approaches – MaxEnt, boosted regression trees, random forests – have approached a performance ceiling, and deep learning has failed to break through on species distribution data. TabPFN, a foundation model that learns to perform Bayesian inference through pretraining on millions of synthetic classification tasks, represents a different paradigm for deep learning: rather than training on each dataset, it applies learned inference patterns to new problems in a single forward pass: it performs 'in-context learning'. We ask whether this new paradigm of learning algorithms can achieve good performance on SDM, despite strong differences in typical SDM datasets relative to TabPFN's training data.

We evaluated TabPFN against established SDM methods across 226 species in six geographic regions using a standardized SDM benchmark. To address the mismatch between TabPFN's pretraining context and the structure of presence-background data, we developed two adaptations: ensemble class balancing, which partitions pseudo-absences across ensemble members while retaining all presence records in each; and domain-specific finetuning through a two-step training process on SDM tasks.

Finetuned TabPFN achieved the highest discrimination on this benchmark, with mean ROC-AUC of 0.762, exceeding MaxNet (0.732), Random Forest (0.727), BRT (0.724), and GAM (0.717) – a 4.1% relative improvement. On held-out species not seen during finetuning, performance was essentially identical (0.763), confirming generalization rather than memorization. Under spatially-separated evaluation, finetuned TabPFN maintained its advantage over all traditional methods (0.699 mean ROC-AUC vs. 0.656-0.683 for traditional models). Using Miller's calibration slope – a ratio calibration measure appropriate for presence-only models – finetuned TabPFN achieved strong probability calibration (slope 1.110), comparable to the best traditional methods.

These results demonstrate that foundation models, when appropriately adapted, can exceed traditional SDM methods. The combination of strong discrimination, sub-second inference, and substantial possibilities for extension, positions TabPFN as a strong alternative for presence-only SDM modeling.

Introduction

Species distribution models (SDMs) have become indispensable tools in ecology and conservation, predicting where species occur based on environmental conditions. These predictions inform decisions that directly affect species survival: identifying critical habitat for endangered taxa, predicting invasion fronts before they establish, prioritizing survey locations for rare species, and assessing vulnerability to climate change [1,2]. As biodiversity loss accelerates and climate shifts redraw the maps of species distributions, the quality of these predictions matters more than ever. A model that better discriminates suitable from unsuitable habitat could mean the difference between a reserve that protects a species and one that misses its actual range by kilometers.

Traditional machine learning methods have dominated SDM for over two decades. Maximum entropy modeling[3,4], boosted regression trees (BRT)[5], random forests (RF)[6], and generalized additive models (GAMs)[7] form the methodological backbone of the field. These approaches handle complex nonlinear relationships between species and their environments while remaining accessible to ecologists without extensive machine learning expertise. MaxEnt, in particular, has become the de facto standard for presence-only modeling, its maximum entropy approach elegantly addressing the challenge of working with occurrence records that lack confirmed absences. Recent comprehensive benchmarking, notably Valavi et al.[8], has systematically compared these methods across hundreds of species and multiple geographic regions. The results confirm their robust performance but also reveal something sobering: despite continuous methodological refinement over the past decade, improvements in predictive accuracy have been incremental rather than transformative. The field appears to have approached a performance ceiling. That is, the best-performing methods in their comparison – despite spanning a range of algorithmic approaches introduced over the past two decades – converged on similar accuracy levels, with differences among top models being small relative to the gap separating them from weaker methods.

This plateau has persisted despite the deep learning revolution that transformed computer vision, natural language processing, and other domains. Deep neural networks have failed to surpass – or even consistently match – traditional methods for SDM. Early applications showed promise but did not deliver substantial gains in predictive performance over established approaches[9], and more recent evaluations confirm the pattern[10,11]. While convolutional networks can capture spatial structure in specific contexts[12], deep learning has not revolutionized SDM as it has in other fields. This apparent failure extends beyond ecology. More generally, across machine learning, researchers consistently observe that deep learning underperforms tree-based methods on small-to-medium tabular datasets[13,14], likely for similar reasons. These reasons include sample sizes typical of ecological data (often hundreds rather than thousands of observations), high correlation among environmental variables, irregular feature importance patterns, and the mixture of continuous and categorical variables common in ecological datasets[15]. For SDM specifically, spatial autocorrelation reduces effective sample sizes while geographic sampling bias introduces systematic errors. These challenges, combined with the need for interpretable predictions that can inform management decisions, have favored traditional methods over deep learning.

Prior-Fitted Networks (PFNs), which learn to perform Bayesian inference by training on millions of synthetic datasets, represent a potential shift in this landscape. PFNs exemplify the emerging paradigm of foundation models – models trained on broad data at scale that can be adapted to a wide range of downstream tasks[16]. Unlike conventional deep learning that requires training on each new dataset, PFNs learn the inference process itself at training time[17]. At inference, these models can make predictions on entirely new datasets in a single forward pass – similar to how large language models solve new tasks from a few examples without retraining. TabPFN[18,19] applies this approach to tabular classification, training a transformer architecture on synthetic datasets generated from structural causal models and Bayesian neural networks. This design encodes a preference for simple causal explanations while maintaining flexibility. The most recent version, TabPFN-2.5[20], handles datasets with up to 50,000 samples and 2,000 features, achieving consistent superiority over previous state-of-the-art (SOTA) tree-based methods while maintaining sub-second inference. It also introduces a distillation engine that can export predictions as interpretable models, potentially addressing concerns about black-box predictions for conservation applications. The theoretical foundations of this approach have been formalized through connections to martingale posteriors, providing rigorous justification for treating TabPFN predictions as Bayesian posterior approximations with well-calibrated uncertainty estimates[21]. This combination of competitive performance, zero hyperparameter tuning, fast inference, and principled uncertainty quantification suggests that foundation models may offer a viable alternative for ecology.

Whether TabPFN can replicate its success in general tabular classification when applied to species distribution modeling remains an open question. SDM presents challenges that may not be well-represented in general benchmarks. Spatial autocorrelation creates dependencies that violate the exchangeability assumptions underlying many machine learning methods, including PFNs. The presence-only data structure common in SDM, where absences are unknown and must be approximated through background sampling, creates extreme class imbalance – ratios of 1:10 to 1:100 are typical – that differs fundamentally from the balanced or moderately imbalanced problems dominating TabPFN's training regime. Most consequentially, SDM models must generalize across geographic space to predict distributions in unsampled regions, a challenge that standard cross-validation fails to assess and that current foundation models may struggle with given their training on datasets without explicit spatial structure.

To evaluate whether this new paradigm can break through the SDM performance ceiling, we tested TabPFN against established SDM methods using the benchmark framework of Valavi et al. (2022), which comprises 226 species distributed across six geographic regions with standardized presence-only training data and independent presence-absence test data. We compared TabPFN against MaxEnt, boosted regression trees, random forest with downsampling, and generalized additive models under both standard evaluation and spatially-blocked evaluation to assess generalization across geographic space. To address the mismatch between TabPFN's pretraining context and the structure of presence-background SDM data, we developed two domain-specific adaptations: ensemble class balancing, which partitions pseudo-absences across ensemble members while retaining all presence records; and two-step finetuning of the TabPFN foundation model on SDM tasks from a broad sample of species.

Methods

2.1 Benchmark Dataset

We evaluated TabPFN using the comprehensive benchmark dataset from Elith et al.[22] – which was thoroughly tested by Valavi et al. (2022)) – and was designed to assess presence-only species distribution modeling methods across diverse biogeographic contexts. The dataset encompasses 226 species distributed across six geographic regions: Australian Wildlife Territory (AWT), New South Wales (NSW), Canada (CAN), New Zealand (NZ), South Africa (SA), and Switzerland (SWI). These regions span substantial variation in spatial extent, environmental gradients, and taxonomic composition. Data were obtained from the disdat R package[22].

For each species, the dataset includes three types of occurrence data: presence/occurrence records (POs) used for model training, background or pseudo-absence points (BGs) representing available habitat, and independent presence-absence test data (PAs) with confirmed presences and absences for evaluation. The distinction between training and test data is crucial: training data follow the presence-only paradigm common in SDM, where absences are unknown and background points substitute as pseudo-absences, while test data provide confirmed presences and absences that enable unbiased performance assessment.

2.1.1 Environmental Covariates

Environmental predictors included climatic variables, topographic features, soil properties, and vegetation characteristics. The specific covariates varied by region, ranging from 7 variables in Canada to 12 in New South Wales and Switzerland:

- **AWT** (8 variables): bc04, bc05, bc06, bc12, bc15, slope, topo, tri
- **CAN** (7 variables): alt, asp2, ontprec, ontslp, onttemp, ontveg, watdist
- **NSW** (12 variables): cti, disturb, mi, rainann, raindq, rugged, soildepth, soilfert, solrad, tempann, topo, vegsys
- **NZ** (11 variables): age, deficit, hillshade, mas, mat, r2pet, slope, sseas, toxicats, tseas, vpd
- **SA** (8 variables): sabio12, sabio15, sabio17, sabio18, sabio2, sabio4, sabio5, sabio6
- **SWI** (12 variables): bcc, calc, ccc, ddeg, nutri, pday, precyy, sfroyy, slope, sradyy, swb, topo

Five variables were treated as categorical factors: vegetation type (ontveg in CAN), vegetation system (vegsys in NSW), toxicity categories and age class (toxicats and age in NZ), and calcareousness (calc in SWI). Four of six regions contain at least one categorical variable.

2.1.2 Data Preprocessing

We converted all occurrence and environmental data to spatial features using the `sf` package in R, matching coordinate reference systems to regional boundaries. For each species, training datasets combined species presences with background points, while test datasets combined presence-absence records. The response variable was coded as a factor with levels "no" and "yes".

The preprocessing pipeline, implemented using `tidymodels` recipes, consisted of three sequential steps: removal of zero-variance numeric predictors, Yeo-Johnson power transformation of all numeric predictors to approximate normality while accommodating zero and negative values, and normalization to mean zero and unit standard deviation. The recipe was fitted on training data alone; the same transformation parameters were then applied to test data to prevent information leakage.

Categorical variables were explicitly converted to factors before preprocessing. For GAMs, categorical variables were retained as factors to enable proper modeling of categorical effects. For TabPFN, categorical variables were handled through explicit specification of categorical feature indices, which were passed to the classifier before fitting. MaxNet, BRT, Random Forest, and GAM all received factor variables as native R factors without conversion. Only the local TabPFN implementation required explicit specification of categorical feature indices, which were passed to the classifier before fitting.

To evaluate spatial generalization, we created spatially-filtered training sets by excluding training points within 10 km of any test location. Standard cross-validation can produce inflated performance estimates by failing to account for spatial sorting bias – the tendency for spatially proximate points to share both environmental and response values[23]. Spatial buffering addresses this bias by ensuring independence between training and test data, following established protocols for spatial blocking in ecological modeling[24,25]. This approach provides a rigorous assessment of model performance in novel geographic areas. Species with single-class datasets after spatial filtering were excluded from spatial evaluation (~41 species), as classification requires representation of both classes.

2.2 Modeling Approaches

2.2.1 Traditional SDM Methods

We compared TabPFN against five established species distribution modeling methods. These methods were selected because they all exceeded the "best models" performance threshold defined in Valavi et al. (2022), which systematically evaluated SDM algorithms across the same benchmark dataset. The traditional methods we tested represent decades of methodological refinement: MaxEnt has been continuously developed since 2004[4], boosted regression trees were adapted for SDM in 2008[5], and random forests and GAMs have similarly long histories of ecological application. For each method, we followed the best-practice configurations specified in Valavi et al. (2022), ensuring that our comparison reflects each algorithm's optimized rather than naive performance. All models used a consistent random seed (32639) for reproducibility.

MaxNet implements Maximum Entropy modeling using generalized linear models with elastic net regularization[3]. MaxEnt, which estimates the distribution that maximizes entropy subject to constraints from observed environmental conditions at presence locations, has become a standard approach for presence-only data. We used the default regularization multiplier (regmult = 1) and included all default feature classes: linear, quadratic, product, threshold, and hinge. Predictions used the complementary log-log link function.

MaxEnt (Java) is the original Java-based MaxEnt implementation accessed via the `dismo` package. This differs from MaxNet in its optimization approach and handling of regularization. We used automatic feature selection and cloglog output format for comparability.

Boosted Regression Trees (BRT) were implemented using `gbm.step()` from the `dismo` package[5], following Valavi et al. (2022). To account for class imbalance, we down-weighted background points relative to presence points by the ratio of presences to backgrounds. Tree complexity was set adaptively: stumps (complexity = 1) for species with fewer than 50 presences, moderate interactions (complexity = 5) for larger samples. We used a learning rate of 0.001, stochastic gradient boosting with 75% bagging fraction, and 5-fold cross-validation to select the optimal number of trees up to a maximum of 10,000.

Random Forest models used the `randomForest` package with balanced sampling[26]. Rather than weighting, we employed per-class downsampling: for each of 1,000 trees, both presence and absence classes were sampled to the size of the minority class. This approach, which directly addresses class imbalance at the tree level, used default parameters for variable selection and node size.

Generalized Additive Models (GAMs) were fitted using the `mgcv` package and using the same region-specific R model formulas used in Valavi et al. (2022), which specified smooth terms for continuous predictors and linear factor terms for categorical variables[27]. Continuous predictors were modeled using penalized thin plate regression splines, while categorical variables were included as factor terms. Models used a binomial family with logit link, restricted maximum likelihood for smoothing parameter estimation, and the same presence-background weighting as BRT.

2.2.2 TabPFN Variants

We tested multiple TabPFN configurations to evaluate the effects of model variant, probability handling, and ensemble strategy on SDM performance. All TabPFN implementations used the Python `tabpfn` package (version 2.5) via R's `reticulate` interface, running on local GPU with CUDA.

TabPFN Default uses the standard TabPFN 2.5 classifier with default parameters: n_estimators = 8, softmax_temperature = 0.9, balance_probabilities = FALSE, average_before_softmax = FALSE. This configuration was pretrained primarily on synthetic classification tasks.

TabPFN Real uses the `v2.5_real` checkpoint, which was additionally finetuned on real-world tabular datasets and may better capture patterns in empirical data distributions.

TabPFN with Balanced Probabilities adds two adjustments to the base configurations. The balance_probabilities setting rescales predicted class probabilities to account for observed class frequencies in the training data. The average_before_softmax setting averages logits across ensemble members before applying the softmax transformation, which tends to produce better-calibrated probability estimates than averaging probabilities directly.

TabPFN Subsample Ensemble (TabPFN-SS) implements a novel SDM-specific ensemble approach described in Section 2.3.

TabPFN Finetuned combines domain-specific finetuning (Section 2.4) with the subsample ensemble approach.

For all TabPFN variants, response variables were converted from factor to numeric (0/1), and categorical feature indices were explicitly specified before fitting. Training used the `fit()` method; predictions used `predict_proba()` to obtain class probabilities. We extracted the probability of presence from the second column of the prediction matrix.

2.3 Ensemble Class Balancing

Species distribution modeling differs fundamentally from typical tabular classification due to the deliberate generation of pseudo-absence data. Standard presence-only SDM workflows create substantial class imbalance – often 1:10 or greater presence:absence ratios – that reflects a data generation strategy rather than true class prevalence. TabPFN was pretrained on balanced or moderately imbalanced tasks, creating a distribution mismatch when applied to SDM data.

The key insight motivating our approach is that presence records and pseudo-absences have fundamentally different epistemic status in SDM. Presence records are valuable and relatively scarce, representing actual observations of species occurrence. Pseudo-absences, by contrast, are interchangeable samples from the background environmental space; any given pseudo-absence point could be replaced by another randomly drawn background point without loss of information. This asymmetry suggests different treatment in ensemble construction.

We developed an ensemble class balancing approach that exploits this asymmetry:

1. Retain all presence records in every ensemble member
2. For each of K ensemble members, draw a balanced random sample of pseudo-absences equal in number to the presences
3. Create K training datasets, each containing all presences and a different balanced subsample of absences, allowing overlap between members
4. Fit TabPFN on each approximately balanced subset
5. Average logits across K ensemble members before applying softmax

This approach uses 100% of both presence data while presenting each ensemble member with approximately balanced classes. The ensemble strategy draws on established principles of deep ensemble methods, which have proven effective for uncertainty estimation and robust prediction in neural networks[28]. We implemented it using TabPFN's native `SUBSAMPLE_SAMPLES` parameter, which enables efficient ensemble creation within a single model call.

The choice of K = 16 ensemble members reflects a balance between computational cost and ensemble diversity. Preliminary experiments (reported in supplementary materials) showed that performance improved with increasing K up to approximately 16 members, with diminishing returns thereafter. The parameter K in our ensemble approach is distinct from TabPFN's internal n_estimators parameter: K determines how many balanced subsets are created from the SDM data, while n_estimators controls the number of model configurations averaged within each TabPFN call. Our subsample ensemble uses n_estimators = 16 with K = 16 balanced subsets, combined with average_before_softmax = TRUE and balance_probabilities = TRUE.

2.4 Domain-Specific Finetuning

While pretrained TabPFN performs competitively on SDM tasks, we hypothesized that domain-specific finetuning could further improve performance by adapting the model to patterns specific to species-environment relationships. SDM tasks differ from TabPFN's pretraining distribution in characteristic ways: environmental predictors exhibit spatial autocorrelation, species-environment relationships often follow unimodal response curves, and the classification boundary typically reflects niche limits rather than arbitrary class separation.

2.4.1 Species Split

To evaluate finetuning on truly held-out species – preventing the model from memorizing benchmark-specific patterns – we split the 226 species into three groups using a fixed random seed:

Finetuning set (62.5%, ~141 species): Used for training during finetuning

Validation set (7.5%, ~17 species): Used for monitoring loss during finetuning

Test set (30%, ~68 species): Held out completely for unbiased evaluation

This species-level split ensures that performance on test species reflects genuine generalization to new taxa, not memorization. The 62.5/7.5/30 split was chosen to provide sufficient training data while maintaining a meaningful held-out test set; the small validation set reflects that validation is used only for early stopping and checkpoint selection, not hyperparameter tuning.

2.4.2 Two-Step Finetuning

We employed a two-step finetuning procedure. Step 1 adapts TabPFN to general SDM classification patterns by training on cross-validation splits within the finetuning species. This step exposes the model to many train/test configurations, encouraging it to learn patterns that generalize across species rather than overfitting to particular species-environment relationships. Step 2 refines the model using the actual benchmark train/test structure, teaching it to perform well under the specific evaluation protocol.

Step 1: Cross-Validation Finetuning. For each epoch, we generate 3-fold cross-validation splits with 5 repeats per species (15 train/test pairs per species). Absences are subsampled to match the presence count in each split. Training uses 40 epochs with learning rate 1e-5 and OneCycleLR scheduling.

Step 2: Benchmark Finetuning. Using the actual train/test pairs from finetuning species, we generate 5 balanced samples per species per epoch. Training uses 100 epochs with learning rate 1e-6, initialized from the best Step 1 checkpoint selected by validation ROC-AUC.

Both steps used PyTorch with CUDA, with validation computed every 900 batches using 16-member ensembles to match inference settings. Maximum training set size was capped at 1,500 samples per split for GPU memory efficiency. Separate finetuned models were trained for non-spatial and spatial evaluation protocols.

2.5 Spatial Evaluation

Standard evaluation used all available training data without spatial considerations, assessing model performance in an "interpolation" setting where test locations may be spatially proximate to training locations. This scenario, while common in SDM benchmarking, may overestimate real-world performance when models are applied to regions without nearby training data.

Spatial evaluation used a 10 km buffer for spatial separation. Training points within 10 km of any test location were excluded, testing model transferability to novel geographic areas. This more stringent evaluation approximates the common conservation scenario of predicting species occurrence in undersampled regions.

Approximately 41 species were excluded from spatial evaluation due to single-class datasets after filtering. These exclusions reflect species with very restricted ranges or limited presence records; their exclusion is necessary for valid classification evaluation but may slightly bias the spatial results toward species with broader distributions.

2.6 Performance Metrics

We evaluated all models using three complementary metrics implemented in the `yardstick` package.

ROC-AUC, which measures the probability that a randomly chosen presence has a higher predicted probability than a randomly chosen absence, serves as our primary discrimination metric. Values range from 0.5 (random) to 1.0 (perfect discrimination). ROC-AUC is threshold-independent and provides a summary of model performance across all possible classification cutoffs.

PR-AUC (Precision-Recall Area Under Curve) provides complementary information for imbalanced datasets. Unlike ROC-AUC, which can remain high even when the model performs poorly on the minority class, PR-AUC directly measures the trade-off between precision and recall and is more sensitive to performance on rare species[29].

Miller's calibration slope and intercept are obtained by regressing observed outcomes on the logit-transformed predicted probabilities[30]. A slope of 1.0 indicates that predicted probability ratios correspond exactly to observed odds ratios – that is, the model's relative confidence is well-scaled. Slopes below 1.0 indicate under-prediction of probability differences between sites, while slopes above 1.0 indicate over-prediction. The intercept captures bias in overall predicted prevalence, though we note that no presence-only model can achieve absolute calibration (intercept = 0) because the baseline prevalence is confounded with the intercept term[31]. We therefore focus on the slope as a measure of ratio calibration, which is identifiable from presence-background data and directly relevant to applications that compare habitat suitability across sites.

We calculated all metrics separately for standard test predictions (generalization), and spatial test predictions (spatial transferability). We report mean values across species as the primary measure of central tendency.

2.7 Held-out versus Independent Test Comparison

To evaluate how well model performance on training-distribution data predicts performance on independent survey data, we constructed holdout test sets from the presence-background training data for each species. Random holdout splits used stratified sampling (80% train, 20% test) with the rsample package. Spatial holdout splits used spatial block cross-validation (spatialsample package, v = 5 folds), retaining the first fold as the holdout test set. Both split types required both classes present in both partitions. We then compared ROC-AUC on these holdout sets – drawn from the same distribution as training data – against ROC-AUC on the independent presence-absence survey data from the Valavi et al. (2022) benchmark. The performance gap (holdout AUC minus independent AUC) quantifies how much a model's apparent performance inflates when evaluated on same-distribution versus truly independent data. We computed this gap for all models, including the 67 test species held out entirely from finetuning, to verify absence of data leakage.

2.8 Computational Environment

All analyses were conducted using R version 4.x with the `targets` package for pipeline management. The pipeline consisted of approximately 7,500 targets representing species-specific models, preprocessing steps, predictions, and metrics.

We used the `crew` package for parallel processing with three controller configurations: 18 workers for traditional SDM models, 1 worker for sequential GPU processing (preventing CUDA memory conflicts), and 2 workers for cloud API requests. Pipeline configuration used transient memory mode and worker-based storage to manage RAM across the large number of targets.

Python environments were managed via `reticulate` with conda environments containing `tabpfn` (version 2.5), `torch`, and `numpy`. Local GPU inference used CUDA with an NVIDIA RTX 4000 ADA. Finetuning was performed using a separate conda environment with PyTorch and CUDA support, with 20-core parallelization for data generation via the `furrr` package.

Code and data are available at `https://github.com/rdinnager/TabPFN-SDM`. Finetuned model checkpoints will be released for reproduction of results.

Results

We evaluated TabPFN against four established species distribution modeling methods – MaxNet, Random Forest, BRT, and GAM – across 226 species from six geographic regions. Additionally, we tested multiple TabPFN configurations including ensemble class balancing variants and domain-specific finetuned models. Models were assessed using three metrics: ROC-AUC and PR-AUC for discrimination performance, and Miller's calibration slope for ratio calibration. Performance was evaluated under two scenarios: non-spatial evaluation using randomly held-out test data (n=226 species), and spatial evaluation using test data separated from training data by a 10 km buffer (n=185 species, with 41 species excluded due to insufficient spatial separation). We report mean values across species as the primary measure of central tendency.

3.1 Overall Model Performance

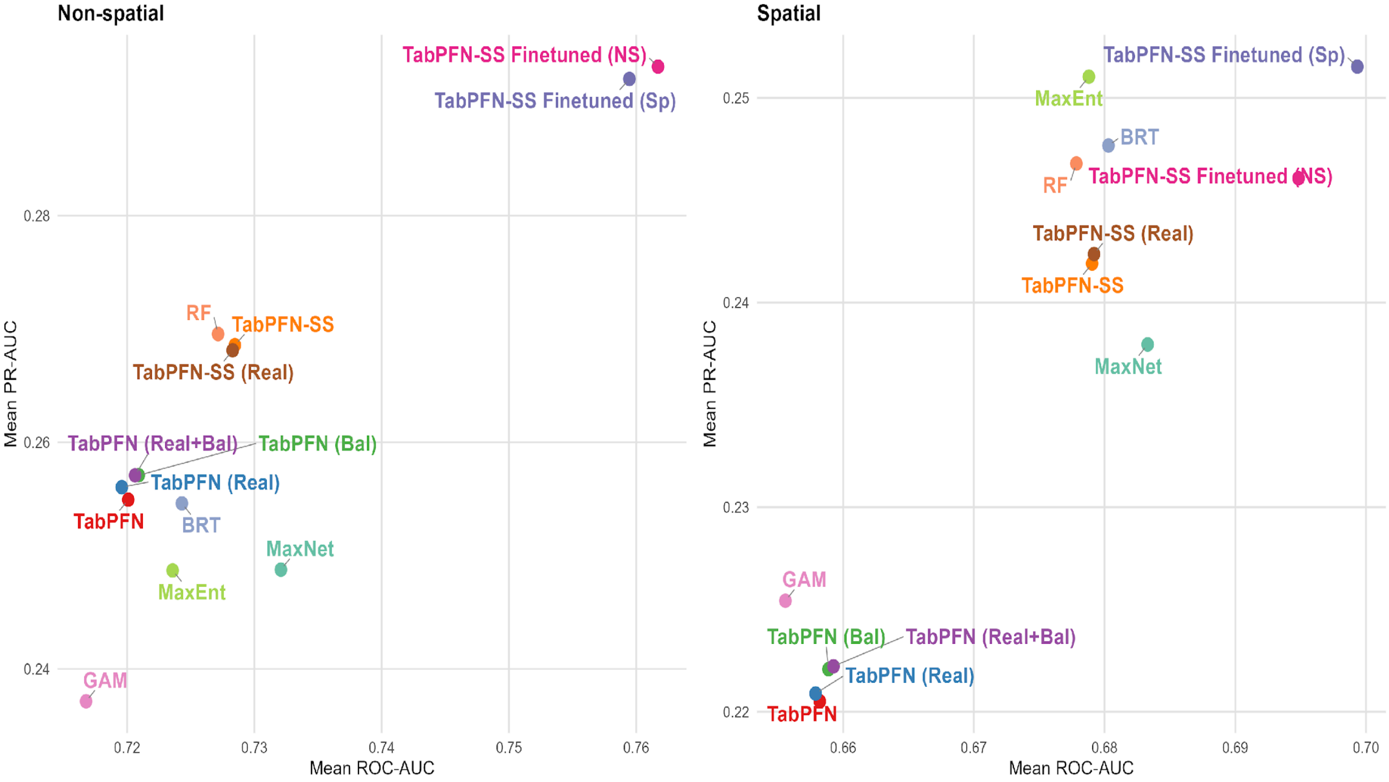

Figure 1. Model performance summary. Mean ROC-AUC (x-axis) versus mean PR-AUC (y-axis) for each model across all species, with non-spatial (left) and spatial (right) evaluation. variants include the original (TabPFN), subsampled ensemble (TabPFN-SS), and finetuned versions. n = 226 species with complete data across all models.

Table 1. Non-spatial model performance. Mean values across 226 species with complete data for all models. Bold values indicate best performance.

Model	**ROC-AUC**	**PR-AUC**	**Miller Slope**	**Miller Intercept**
TabPFN-SS Finetuned (Non-spatial)	**0.762**	**0.293**	1.110	-8.405
TabPFN-SS Finetuned (Spatial)	**0.759**	**0.292**	1.295	-9.718
TabPFN-SS (Default)	0.728	0.269	0.679	-6.294
TabPFN-SS (Real)	0.728	0.268	0.689	-6.362
TabPFN Default+Avg+Bal	0.721	0.257	0.501	-2.655
TabPFN Real+Avg+Bal	0.721	0.257	0.501	-2.674
TabPFN Default	0.720	0.255	0.507	**-0.201**
TabPFN Real	0.720	0.256	0.509	-0.215
MaxNet	0.732	0.249	0.386	-2.258

Random Forest	0.727	0.270	0.735	-2.500
BRT	0.724	0.255	**0.920**	-2.682
MaxEnt	0.724	0.249	0.369	-2.245
GAM	0.717	0.237	0.444	-2.654

Across all species and regions, the finetuned TabPFN models achieved the highest discrimination performance among all methods tested (Figure 1; Table 1). In non-spatial evaluation, the finetuned TabPFN achieved a mean ROC-AUC of 0.762, followed closely by the spatially-finetuned variant at 0.759. The subsample ensemble TabPFN variants without finetuning achieved mean ROC-AUC of 0.728, while standard local TabPFN configurations ranged from 0.720 to 0.721.

Among traditional methods, MaxNet had the highest mean ROC-AUC (0.732), followed by Random Forest (0.727), BRT and MaxEnt (both 0.724) and GAM (0.717). The finetuned TabPFN thus exceeded the best traditional methods by 0.035 in mean ROC-AUC, a 4.8% relative improvement.

Precision-recall analysis showed similar patterns. The finetuned TabPFN achieved the highest mean PR-AUC (0.293), followed by the spatially-finetuned variant (0.292) and subsample ensemble variants (0.268-0.269). Among traditional methods, Random Forest achieved the highest mean PR-AUC (0.270), followed by BRT (0.255), MaxEnt (0.249), MaxNet (0.249), and GAM (0.237). The PR-AUC results are particularly relevant for applications involving rare species, where the high imbalance between presences and absences makes precision a more informative metric than overall accuracy..

The circular parallel coordinates plots (Figures 2a, 2b) reveal per-species consistency and variation across models. Most species showed similar patterns across methods, with relatively few species exhibiting strong model-specific preferences. Regional variation in model performance exceeded between-model variation for most comparisons – a finding that suggests geographic and ecological factors exert substantial influence on predictive performance, perhaps more so than method choice.

3.2 Effect of Ensemble Class Balancing

The subsample ensemble approach, which retains all presence records while drawing balanced subsamples of pseudo-absences for each ensemble member, improved TabPFN performance over standard configurations (Figure 5a).

In non-spatial evaluation, the subsample ensemble variants achieved mean ROC-AUC of 0.728 compared to 0.720–0.721 for standard local TabPFN configurations, representing a 1.0–1.1% improvement. This improvement held consistently across regions and metrics. The approach addresses a fundamental mismatch between TabPFN's pretraining distribution–which assumes moderate class imbalance–and the extreme imbalance characteristic of presence-background SDM data (often 1:10 to 1:100).

The ensemble approach also improved calibration, as measured by Miller's calibration slope (ideal value = 1.0). The TabPFN-SS variants achieved mean Miller slopes of 0.68–0.69, closer to ideal than standard local TabPFN configurations (0.50–0.51), indicating that the subsampling ensemble partly corrects the probability compression typical of TabPFN on imbalanced data. However, BRT achieved the best calibration among all methods (mean slope 0.92), followed by Random Forest (0.74). MaxEnt (0.37), MaxNet (0.39), and GAM (0.44) showed the most compressed probability scales

The subsample ensemble approach enabled memory-efficient training on consumer GPUs by reducing the effective batch size while maintaining 100% data utilization. This memory efficiency proved essential for enabling the domain-specific finetuning described below.

3.3 Effect of Domain-Specific Finetuning

Domain-specific finetuning using a two-step training process yielded substantial performance gains across all metrics (Table 1, Table 2; Figures 3, 4).

In non-spatial evaluation, finetuned models achieved mean ROC-AUC of 0.762 (non-spatial finetuning) and 0.759 (spatial finetuning), with corresponding PR-AUC of 0.293 and 0.292 respectively. Both finetuned variants exceeded all other methods, including the best traditional method (MaxNet, 0.732), by 0.027–0.030 in mean ROC-AUC.

The critical test of finetuning validity is performance on species not seen during training. Performance on the 67 held-out test species (30% not seen during finetuning) confirmed that finetuning generalizes to unseen species (Table 2). On held-out species, the finetuned models achieved mean ROC-AUC of 0.763 for both non-spatial and spatial finetuning variants. These held-out species results were essentially identical to the full dataset results (0.762-0.763), providing strong evidence against overfitting to the training species.

Comparing finetuned models to the subsample ensemble baseline on held-out test species, 53.7% of species showed improvement in ROC-AUC, 52.2% in PR-AUC, and 59.7% in calibration (finetuned non-spatial vs TabPFN-SS). Comparing to standard TabPFN configurations showed larger improvements: 61.2% of species improved in ROC-AUC and 65.7% in PR-AUC, with mean improvement magnitude of +0.018 ROC-AUC and +0.034 PR-AUC.

On the other hand, these percentages indicate that 38.8-46.3% of species did not improve or degraded with finetuning. We did not identify systematic characteristics distinguishing improved from non-improved species in this analysis, though such patterns warrant further investigation.

3.4 Calibration Performance

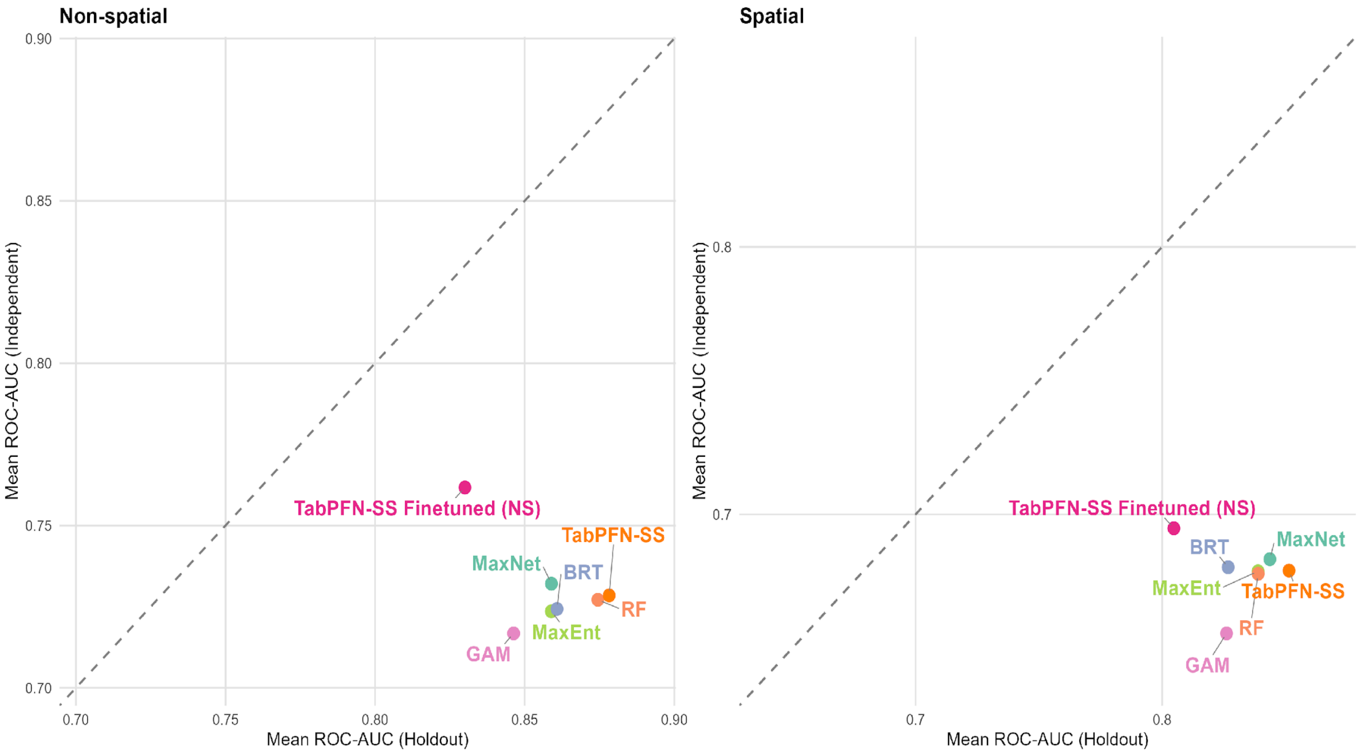

Figure 2. Comparison of held-out versus independent test discrimination. Mean ROC-AUC on held-out test samples (x-axis) versus mean ROC-AUC on independent test samples (y-axis) for each model, with non-spatial (left) and spatial (right) evaluation. The dashed line indicates the 1:1 relationship; points below the line indicate performance degradation on truly independent samples.

In non-spatial evaluation, BRT achieved the mean Miller's slope closest to 1.0 at 0.920 (deviation of 0.080 from ideal), making it the best-calibrated model by this metric. The finetuned TabPFN (non-spatial) was the second-closest at 1.110 (deviation of 0.110), and notably the only model to exceed 1.0, indicating slight over-prediction of occurrence probability ratios. Random Forest achieved a slope of 0.735, while the subsample ensemble TabPFN variants ranged from 0.679 to 0.689 and the local TabPFN variants from 0.501 to 0.509. Among the remaining traditional methods, GAM produced a slope of 0.444, MaxNet 0.386, and MaxEnt 0.369 – all substantially below 1.0, indicating systematic under-prediction of probability ratios.

Under spatial evaluation, calibration patterns shifted substantially. BRT maintained the slope closest to 1.0 at 0.854 (deviation of 0.146), followed closely by the subsample ensemble TabPFN variants (0.823–0.848) and Random Forest (0.791). In contrast, the finetuned TabPFN variants showed dramatically inflated slopes of 1.896 (non-spatial finetuned) and 2.166 (spatially finetuned), indicating severe over-prediction of occurrence probability ratios – a striking reversal from their strong non-spatial calibration. GAM produced a slope of −3.318, indicating catastrophic miscalibration under spatial evaluation. These results suggest that while finetuning improved discrimination (ROC-AUC), it came at a substantial cost to ratio calibration under spatial evaluation, whereas the subsample ensemble approach preserved calibration across both evaluation conditions.

3.5 Held-out versus Independent Test Performance

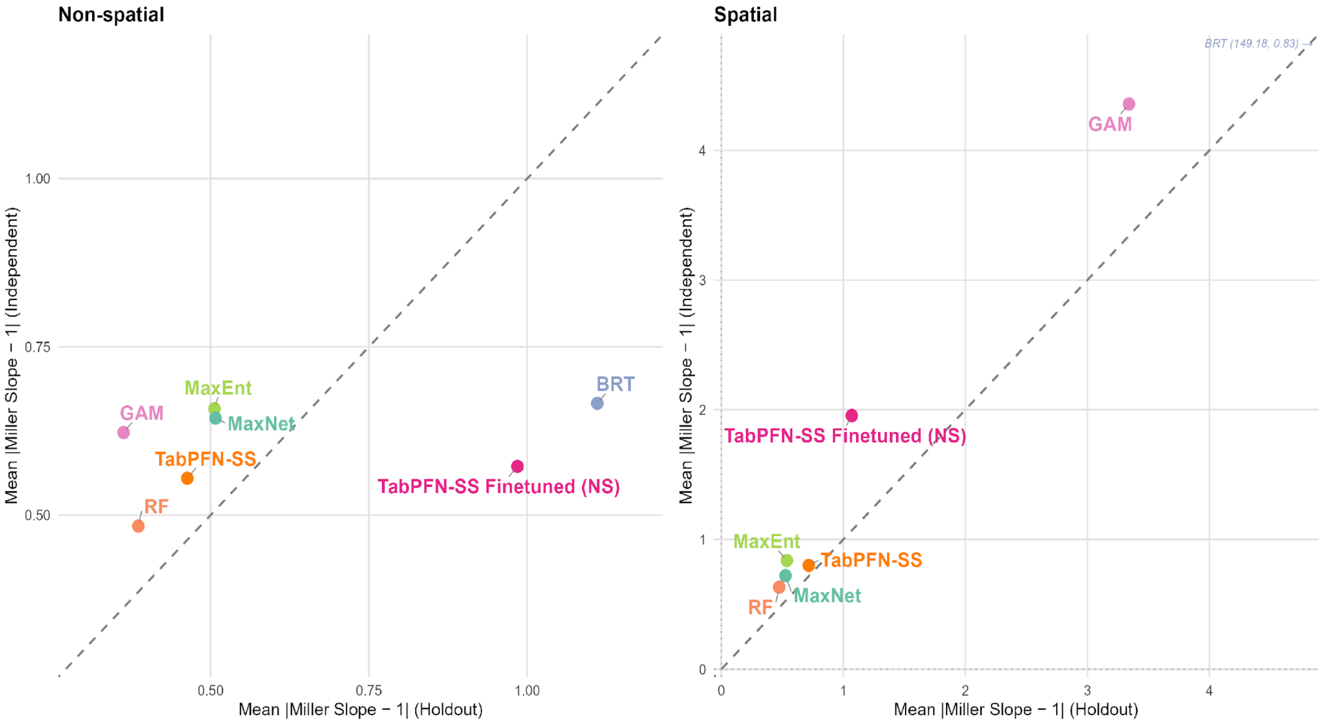

Figure 3. Comparison of held-out versus independent test calibration deviation. Mean calibration deviation (|Miller slope − 1|) on held-out samples (x-axis) versus independent samples (y-axis) for each model, with non-spatial (left) and spatial (right) evaluation. Values closer to zero indicate better calibration. The dashed diagonal indicates equal performance on both evaluation sets; points below the diagonal indicate better calibration on independent data. Dotted lines mark ideal calibration (zero deviation).

All models performed better on holdout data drawn from the training distribution than on independent presence-absence survey data, but the magnitude of this gap varied substantially across methods. In non-spatial evaluation, the finetuned TabPFN (non-spatial) showed the smallest performance gap at 0.068 ROC-AUC units. Default TabPFN had a gap of 0.150 – more than double – while Random Forest (0.147), BRT

(~0.14), MaxNet (~0.13), and GAM (~0.13) fell between these extremes. Finetuning reduced the gap by 54.4% compared to the default TabPFN configuration.

The 67 held-out species not seen during finetuning confirmed that this smaller gap does not reflect data leakage. These species achieved an independent ROC-AUC of 0.763 with a holdout-to-independent gap of just 0.056 – even smaller than the gap across all species. If finetuning had overfit to the benchmark structure, held-out species would show a larger gap, not a smaller one.

3.6 Spatial Generalization

Performance declined across all models when evaluated on spatially separated test data (n=185 species), though the magnitude of decline was more similar across methods than median values suggested.

In spatial evaluation, the finetuned TabPFN variants achieved the highest mean ROC-AUC (0.695-0.699), followed by MaxNet (0.683), BRT (0.680), TabPFN-SS and MaxEnt (0.679), Random Forest (0.678), and GAM (0.656). The performance drop from non-spatial to spatial evaluation quantifies each method's spatial transferability: TabPFN-SS showed a 6.7% drop (0.728 to 0.679), finetuned models showed 8.3–8.4% drops (0.762 to 0.699 or 0.759 to 0.695), BRT showed a 6.1% drop (0.724 to 0.680), Random Forest showed a 6.7% drop (0.727 to 0.678), and MaxNet showed a 6.7% drop (0.732 to 0.683)..

The performance drops were thus comparable across methods, ranging from 6.1% to 8.6%. TabPFN variants showed somewhat larger drops than traditional methods, but the difference was modest. The finetuned TabPFN maintained its discrimination advantage even under spatial evaluation, achieving the highest absolute performance despite the slightly larger relative drop.

Table 2. Spatial model performance. Mean values across 185 species with complete data for all models. Bold values indicate best performance.

Model	ROC-AUC	PR-AUC	Miller Slope	Miller Intercept
TabPFN-SS Finetuned (Spatial)	**0.699**	**0.252**	2.166	-16.980
TabPFN-SS Finetuned (Non-spatial)	**0.695**	0.246	1.896	-15.068
TabPFN-SS (Real)	0.679	0.242	0.848	-7.842
TabPFN-SS (Default)	0.679	0.242	0.823	-7.684
TabPFN Real+Avg+Bal	0.659	0.222	0.440	-2.747
TabPFN Default+Avg+Bal	0.659	0.222	0.435	-2.745
TabPFN Default	0.658	0.221	0.441	**-0.332**
TabPFN Real	0.658	0.221	0.439	-0.344
MaxNet	0.683	0.238	0.330	-2.248
BRT	0.680	**0.248**	**0.854**	-2.384
MaxEnt	0.679	**0.251**	0.461	-2.321
Random Forest	0.678	**0.247**	0.791	-2.368
GAM	0.656	0.225	-3.318	-59.788

The spatially-finetuned model showed improved spatial generalization compared to non-spatial finetuning on the held-out test species (Table 2). On spatial test data, the spatially-finetuned model achieved a mean ROC-AUC of 0.717 compared to 0.707 for the non-spatially-finetuned model. Among held-out species, 62.5%

showed improved spatial ROC-AUC with spatial finetuning compared to standard TabPFN, with a mean improvement of +0.036.

3.7 Regional Performance Patterns

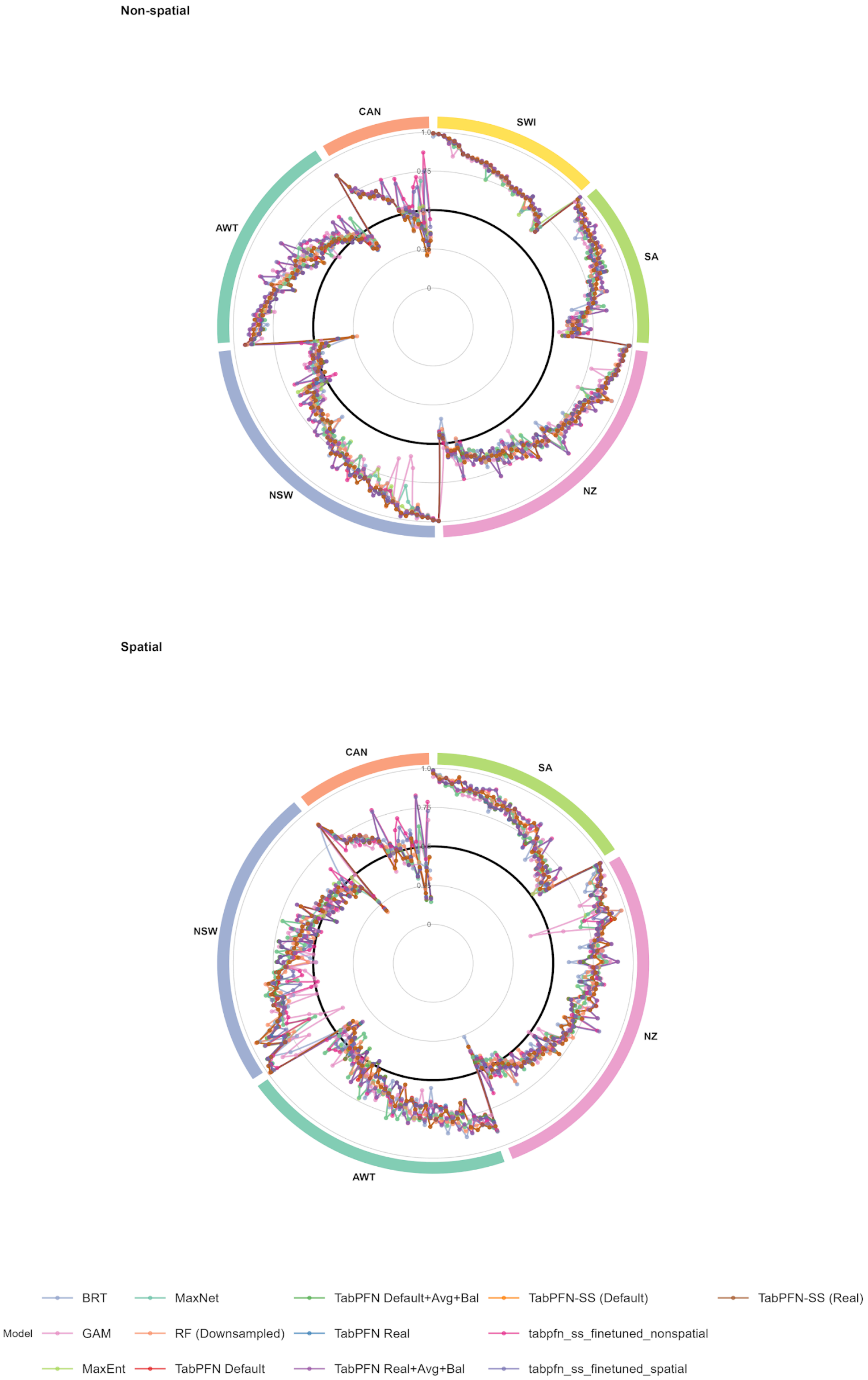

Figure 4. Species-level model performance across six geographic regions. Circular plot showing ROC-AUC for individual species arranged by region: AWT (Australian Wildlife), CAN (Canada), NSW (New South Wales), NZ (New Zealand), SA (South Africa), and SWI (Switzerland). Radial distance from centre indicates ROC-AUC (0.0 to 1.0). Colored lines represent different models, revealing both species-specific variation and regional patterns in model performance. The black circle marks ROC-AUC = 0.5 (random classifier baseline).

Model performance varied substantially across the six geographic regions (Table 2; Supplementary Tables S2-S3).

South Africa and Switzerland emerged as the best-performing regions, with finetuned TabPFN achieving mean ROC-AUC of 0.829-0.830 in South Africa and BRT achieving 0.822 in Switzerland (finetuned TabPFN reached 0.824). New Zealand showed intermediate performance, with finetuned TabPFN achieving mean ROC-AUC of 0.770.

Canada proved the most challenging region. All models showed reduced performance, with finetuned TabPFN achieving the best mean ROC-AUC (0.634-0.663), compared to MaxNet (0.593), GAM (0.589), and BRT (0.575). The difficulty appears consistent across methods and likely reflects data characteristics or environmental complexity specific to this region.

The finetuned TabPFN variants ranked first or second in all 6 regions for non-spatial evaluation. The finetuning benefit was most pronounced in Canada, where finetuned TabPFN exceeded the next-best method (MaxNet) by 0.040 in mean ROC-AUC – a larger margin than observed in any other region.

That regional variation exceeded model variation across most comparisons is perhaps the most striking finding of this analysis. It suggests that understanding what makes certain regions challenging may be more consequential for improving SDM predictions than further method development.

3.8 Computational Efficiency

Computational efficiency varied substantially across methods (Figure 5). MaxEnt was the fastest method with a median fit+predict time of 1.90 seconds per species. Random Forest followed at 2.31 seconds, then the TabPFN-SS variants at 6.89–8.41 seconds.

Standard local TabPFN configurations required 13.36–13.62 seconds median time. MaxNet had a median of 7.67 seconds but with extremely high variance (sd = 132 seconds, max = 1,820 seconds) due to convergence issues on some species. BRT required substantially more time (median 77.92 seconds), while GAM had a lower median (43.39 seconds) but enormous variance (sd = 232 seconds, max = 1,871 seconds), making it the most unpredictable method; by mean computation time, BRT and GAM were nearly identical (~114 seconds).

For the majority of species, TabPFN-SS provided rapid inference under 15 seconds, making it practical for large-scale applications. The mean computation time for TabPFN-SS (10.94–13.78 seconds) was moderately higher than the median, reflecting occasional slower runs but no extreme outliers. MaxNet, by contrast, showed occasional extreme computation times (max 1,820 seconds), which could affect workflow planning for batch processing.

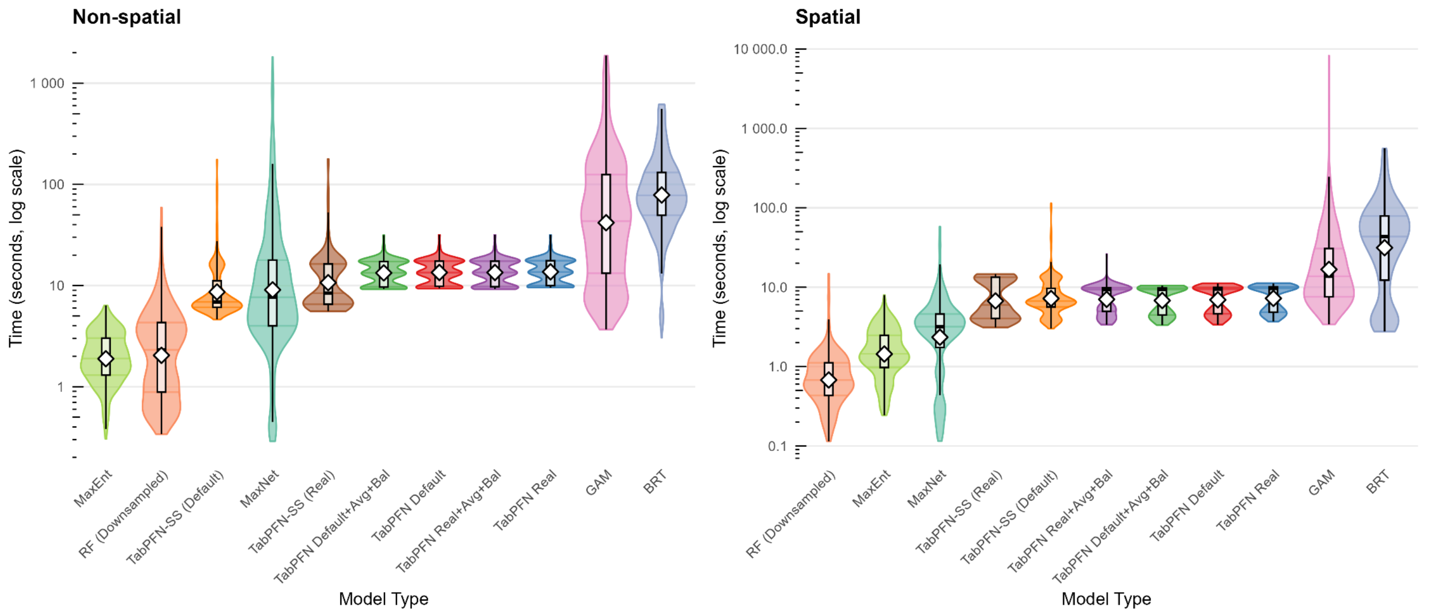

Figure 5. Distribution of computation time (fit + predict) per species for each model, under non-spatial (left) and spatial (right) evaluation. Violins show the density distribution; boxes indicate the interquartile range; white diamonds mark the mean. The y-axis is log-scaled. Models are ordered by median computation time. Finetuned TabPFN variants are not shown separately because finetuning modifies only the model weights, not the inference architecture; their computation times are identical to the corresponding TabPFN-SS variants.

3.9 Summary

The finetuned TabPFN models achieved the highest discrimination performance across all methods tested, with mean ROC-AUC of 0.762 compared to 0.732 for MaxNet, the best traditional method. The ensemble class balancing approach improved pretrained TabPFN performance from 0.720 to 0.728 mean ROC-AUC, while domain-specific finetuning provided an additional 0.034 improvement. On held-out species not seen during finetuning, 61.2% showed improved ROC-AUC, confirming generalization. Held-out species achieved essentially identical mean performance (0.763) to the full dataset (0.762), providing strong evidence against overfitting.

Discussion

TabPFN, a foundation model for tabular classification, achieves the best overall – or "state-of-the-art" (SOTA) in machine learning terminology – species distribution modeling performance when appropriately adapted to the SDM context. Across the benchmark species, finetuned TabPFN achieves the highest discrimination among all methods tested, exceeding all traditional methods. This result rests on two methodological contributions that address the mismatch between TabPFN's pretraining context and the distinctive structure of presence-background SDM data.

We note at the outset that our evaluation focuses on prediction – the ability to correctly classify sites as occupied or unoccupied – rather than inference about the ecological mechanisms driving species distributions. Prediction and inference are distinct goals that may require different methods and evaluation criteria. Calibration, which we assess via Miller’s slope, bridges these goals: well-calibrated predictions can inform probabilistic inference about relative habitat suitability, even when the underlying causal processes are not directly modeled.

The first contribution, ensemble class balancing, addresses the extreme class imbalance inherent to presence-background modeling. TabPFN was trained on synthetic datasets with balanced or moderately imbalanced classes, while SDM practice deliberately creates ratios of 1:10 to 1:100 through pseudo-absence sampling. Our approach retains all presence records, which are valuable and relatively rare, while partitioning pseudo-absences across ensemble members so that each member sees approximately balanced classes. This strategy lifts default TabPFN to competitive performance without any domain-specific training.

The second contribution, domain-specific finetuning, pushes performance beyond traditional methods. On held-out test species not seen during finetuning, a majority showed improvement across different comparison baselines. That held-out species achieved essentially identical mean performance to the full dataset confirms that finetuning captures general SDM patterns rather than memorizing individual species.

These contributions compound: the memory efficiency gained from ensemble class balancing enabled finetuning on consumer-grade GPU hardware (NVIDIA RTX 4060, 8GB VRAM). Without this reduction in memory requirements, the experiments producing SOTA results would not have been feasible on accessible hardware. For practitioners seeking to apply TabPFN to species distribution modeling, we recommend the subsample ensemble configuration (TabPFN-SS) as the default starting point, with domain-specific finetuning where GPU resources permit.

Comparison with Previous Work

These results address a persistent question in ecological machine learning: whether deep learning can match or exceed traditional methods for tabular ecological data. Previous work consistently found that deep neural networks underperform tree-based methods on small-to-medium tabular datasets (Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022), including in SDM applications specifically (Kellenberger et al., 2024).

TabPFN represents a fundamentally different paradigm. Rather than training a network from scratch on each dataset, TabPFN learns general patterns of tabular structure during pretraining on millions of synthetic datasets, then applies this knowledge through in-context learning. The architecture learns to perform Bayesian inference implicitly (Deng et al., 2024), explaining its strong performance with minimal data. Our results with domain-adapted TabPFN provide the first demonstration, to our knowledge, that this paradigm can exceed traditional SDM methods when appropriately configured.

The Valavi et al. (2022) benchmark framework enables direct comparison with previous SDM evaluations, where ensemble methods – especially BRT and Random Forest – achieved the best performance. That fine-tuned TabPFN exceeds these benchmarks suggests foundation model approaches represent a new frontier for SDM methodology.

The Magnitude of Improvement in Historical Context

A 0.035 improvement in mean ROC-AUC may appear modest in absolute terms, but context reveals its significance. The traditional methods we tested – MaxEnt, boosted regression trees, random forest, and GAMs – were not selected arbitrarily. Each exceeded the "best models" performance threshold in Valavi et al. (2022), and each represents the endpoint of years to decades of methodological refinement. MaxEnt has been continuously developed and optimized for SDM since 2004 (Phillips et al., 2006, 2017); boosted regression trees were specifically adapted for ecological applications in 2008 (Elith et al., 2008); random forests and GAMs have similarly long histories of tuning for presence-background data. We implemented each method following the best-practice configurations from Valavi et al. (2022)[8], ensuring our comparison reflects optimized rather than naive implementations.

Despite this history of refinement, the top traditional methods cluster tightly in performance. MaxNet has the best performance at 0.732 mean ROC-AUC, with random forest closely behind at 0.727; BRT and MaxEnt achieve 0.724; GAM reaches 0.717. The entire spread among these well-established methods is just 0.015 ROC-AUC units. The gap between the top tier (MaxNet/random forest) and the second tier (BRT/MaxEnt) is even smaller: 0.005. This convergence is not coincidental – it reflects decades of optimization pushing these algorithms toward similar performance ceilings on the same fundamental problem.

Finetuned TabPFN's improvement of 0.030 over the best traditional method is 2 times larger than the entire spread among traditional methods (0.015), and roughly 6 times larger than the gap separating the top traditional performers (0.005). In other words, the jump from traditional methods to finetuned TabPFN exceeds the cumulative differentiation achieved by decades of methodological refinement across multiple algorithmic families. This margin suggests that the foundation model paradigm – learning general inference patterns from synthetic data, then adapting to domain-specific applications – may offer a path beyond the performance plateau that has characterized SDM methodology in recent years.

We do not claim that these ROC-AUC differences translate directly to conservation outcomes; the biological significance of discrimination improvements remains largely unexplored. But from a purely methodological standpoint, an improvement that exceeds the spread among all top-performing traditional methods represents a meaningful advance in a field where incremental gains have been the norm.

SDM as Density Ratio Estimation

An underappreciated connection in the ecological literature links presence-background SDM to density ratio estimation, a framework with deep roots in machine learning[32]. When we model presence locations against background samples, we implicitly estimate the ratio f_presence(e) / f_background(e) – the relative density of species occurrences compared to available environmental space[33,34]. MaxEnt, the most widely used SDM method, can be derived directly from this density ratio framework.

This framing connects SDM to noise contrastive estimation[35] and illuminates why class imbalance matters for TabPFN. In density ratio estimation, the ratio of positive to negative samples determines the normalizing constant; extreme imbalance distorts this relationship. Our ensemble class balancing approach effectively recalibrates this ratio for each ensemble member.

The connection has practical implications beyond TabPFN. Density ratio estimation appears across diverse ML applications – covariate shift adaptation, mutual information estimation, generative modeling, anomaly detection – and the challenges we address here may translate to those contexts. Conversely, the density ratio literature offers methodological insights that ecologists have largely developed in isolation. By highlighting this bridge, we hope to facilitate exchange between communities: our ensemble approach may prove useful for density ratio problems beyond ecology, while advances in that literature may improve SDM methodology. This theoretical grounding suggests that the ensemble class balancing approach we develop here may benefit density ratio problems beyond ecology, just as SDM may benefit from methodological advances in that broader literature.

Spatial Generalization

All methods show performance drops under spatial evaluation that are of comparable magnitude. TabPFN maintains its discrimination advantage under spatial evaluation, achieving the highest performance among all methods.

Finetuning on spatially-separated data improves spatial generalization – a majority of species showed improvement compared to default TabPFN. The spatial-finetuned model exceeds all traditional methods. For applications requiring spatial extrapolation, which includes most applied SDM work, finetuned TabPFN appears to be a strong choice.

This finding aligns with the broader SDM literature, where spatial transferability challenges all methods[36,37]. The relatively similar performance drops across methods suggest that spatial generalization limitations are fundamental to the SDM problem rather than method-specific.

Ratio Calibration

Calibration – whether predicted probabilities reflect actual occurrence rates – matters for any SDM application that uses predictions as more than a ranked list. For presence-only models, however, absolute calibration (the intercept) is unidentifiable because true prevalence is unknown. What we can assess is ratio calibration: whether the relative differences in predicted probabilities correspond to actual differences in occurrence odds.

Miller's calibration slope[30] tests exactly this property by regressing observed outcomes on the logit of predicted probabilities. A slope of 1 means that a site predicted to be twice as suitable genuinely has twice the odds of occurrence. Van Calster et al. (2016)[38] placed this ratio calibration within a formal calibration hierarchy, demonstrating that it is the strongest calibration property testable for presence-only models where true prevalence is unknown[31]. In our benchmark, finetuned TabPFN achieved a Miller slope of 1.110 in non-spatial evaluation (mean across 226 species), indicating slight over-prediction of probability contrasts but overall strong ratio calibration. BRT achieved 0.920, indicating slight under-prediction. All methods showed slopes below 1 except finetuned TabPFN, which slightly exceeded it. Under spatial evaluation, calibration slopes decreased for all methods, consistent with the general performance degradation when predicting across geographic space.

This has practical consequences. When predictions are used to compare habitat suitability across sites – as in reserve design, habitat prioritization, or connectivity modeling – ratio calibration determines whether the predicted differences can be taken at face value. A model with slope 0.386 (MaxNet) predicts a 39% change in log-odds where the true change is 100%; a model with slope 1.110 (finetuned TabPFN) predicts 111% – a slight overshoot but substantially closer to exact. For applications that require only ranking sites, discrimination suffices and any of the tested methods performs adequately. For applications that compare magnitudes of suitability differences, ratio calibration becomes the relevant property.

We note that no presence-only model can achieve absolute calibration (intercept = 0) because the baseline prevalence is confounded with the intercept term[31]. The large negative intercepts across all models reflect this fundamental identifiability constraint rather than model failure. What distinguishes the models is not their intercepts but their slopes – whether doubling the predicted probability corresponds to doubling the actual odds. For practitioners, ratio calibration should be assessed alongside discrimination when selecting an SDM method, particularly for applications that compare magnitudes of habitat suitability across sites.

Generalization to Independent Data: The Role of Pretrained Representations

The gap between holdout and independent test performance measures how much a model has absorbed dataset-specific biases rather than genuine ecological signal. Training presence data contain sampling artifacts – spatial clustering near roads and population centers, taxonomic biases toward charismatic species, habitat accessibility gradients – that correlate with outcomes in holdout data drawn from the same biased sample but not in independently collected surveys. All models showed higher holdout than independent ROC-AUC, but the magnitude of this inflation varied substantially: from 0.069 (finetuned TabPFN) to 0.150 (default TabPFN).

Models trained from scratch on biased data have no mechanism to distinguish ecological signal from sampling artifact. They optimize on whatever features best separate presences from pseudo-absences, and if spatial clustering or road proximity is predictive, they learn it. Geirhos et al. (2020)[39] termed this "shortcut learning" – the tendency of models to exploit the easiest discriminative features, which are often dataset-specific rather than domain-general. Random Forest, BRT, MaxNet, and GAM all exhibit this pattern.

Finetuned TabPFN operates differently. Its pretrained representations, learned from millions of synthetic tabular datasets, encode general patterns of how features relate to binary outcomes[16]. Finetuning adapts these representations to the SDM domain, but the pretrained weights act as an implicit regularizer – they resist wholesale absorption of dataset-specific patterns because the model already has strong priors about what constitutes a general classification signal. The finetuned model's gap of 0.069 represents a 54% reduction compared to default TabPFN (0.150), and the 67 held-out species show an even smaller gap of 0.056, confirming that this transfer ability is not explained by data leakage.

This capacity for cross-dataset learning is precisely what distinguishes foundation models from classical machine learning. Random Forest, BRT, MaxNet, and GAM train independently on each dataset, discarding everything learned when moving to the next species. Their architectures lack the mechanisms for meta-learning – there is no shared representation that accumulates knowledge across tasks. One form of cross-dataset optimization that is possible for classical methods is hyperparameter tuning: selecting algorithm settings that generalize across species rather than optimizing for each one independently. We tested this by tuning Random Forest hyperparameters (mtry, minimum node size, maximum nodes, and number of trees) on the same 62.5% of training species used for TabPFN finetuning, then evaluating on the 30% held-out test species. Tuned RF achieved mean ROC-AUC of 0.753 on held-out species in non-spatial evaluation, compared to 0.756 for default RF – no improvement. Under spatial evaluation, the pattern was identical (tuned: 0.697 vs. default: 0.700). Cross-dataset hyperparameter tuning, the only avenue available to classical methods, produced no measurable benefit for SDM, reinforcing that the gains from TabPFN finetuning reflect genuine representational learning rather than simple parameter optimization.

Regional Variation Exceeds Model Variation

A pattern recurring across our results is that regional variation in performance exceeds model variation. For most species, the choice of geographic region affects predicted performance more than the choice of algorithm. Canada proved the most challenging region, with all models showing reduced performance. The difficulty appears consistent across methods and likely reflects data characteristics or environmental complexity specific to this region.

This finding reinforces that data quality, environmental gradient representation, and region-specific ecological factors are primary determinants of SDM success. The 3-5% ROC-AUC difference between methods is smaller than the 15-25% difference between regions. Practitioners can often choose models based on secondary considerations – computational resources, interpretability requirements, or ease of implementation – without substantial predictive penalty.

Two regions in this benchmark – AWT and NSW – include multiple taxonomic groups, providing a limited window into how model performance varies across taxa. In the Australian Wet Tropics, birds (20 species, median ~146 presences) and plants (20 species, median ~32 presences) achieved broadly similar discrimination despite the substantial difference in sample size. In New South Wales, performance varied more markedly across eight taxa groups: groups with fewer occurrence records (e.g. rainforest trees, bats) consistently yielded lower discrimination regardless of modelling method, while better-sampled groups (e.g. diurnal birds, open-forest trees) achieved higher scores across all models. These patterns are unsurprising – rare species with few records are inherently harder to model – but they underscore that variation among species and taxa often exceeds variation among modelling methods. Whether TabPFN offers any differential advantage for data-sparse taxa is a question the anonymised benchmark species cannot fully resolve, but it merits investigation with attributed occurrence records.

TabPFN’s Cross-Domain Adaptability

TabPFN’s strong performance in species distribution modelling is not an isolated finding. The same model has been adapted to new types of data by adapting how data is prepared for the model, rather than redesigning the model architecture itself. Eremeev et al. (2025)[40] converted graph node classification into tabular problems by aggregating neighbourhood features, then applied TabPFN directly; the resulting system matched well-tuned graph neural networks without finetuning and surpassed them after adaptation. Hoo et al. (2025)[41] took a similar approach to time series forecasting, encoding temporal structure via running indices, calendar features, and spectral decomposition. Despite being roughly 100 times smaller than purpose-built forecasting models such as TimesFM-2.0, TabPFN achieved competitive accuracy across standard benchmarks.

A common pattern connects these applications with our SDM results: domain-specific data is translated into tabular form through relatively straightforward feature engineering, and TabPFN’s in-context learning handles the inference. In SDM, environmental covariates and species occurrence records already constitute tabular data, which may partly explain why TabPFN required no architectural modification to outperform established methods. The model appears to function as a general-purpose learned kernel that adapts its inference strategy to whatever tabular structure it encounters.

Zheng et al. (2025)[42] offer a theoretical account that is consistent with this interpretation. Their analysis demonstrates that TabPFN exhibits what they term “spectral adaptivity” – the model automatically adjusts its effective bandwidth to match the density and complexity of the input data. As sample count increases, the context kernel spectrum flattens, allowing finer-grained pattern capture. This property may explain the adaptability we observe across SDM species with varying range sizes and habitat specificities: rather than requiring manual hyperparameter tuning to match different levels of spatial complexity, TabPFN adjusts its inference granularity through the data itself. The common thread across these applications is that domain adaptation of TabPFN relies on thoughtful data preparation rather than model re-engineering, lowering the barrier for researchers in new fields.

Limitations

Several limitations constrain interpretation of these findings.

Our results derive from the Valavi et al. (2022) benchmark, which, though comprehensive (226 species, six regions, independent presence-absence test data), may not generalize to all SDM applications. Presence-only models are susceptible to sample selection bias – the tendency for occurrence records to be systematically biased toward accessible or heavily sampled areas[43] – which our benchmark data may reflect despite efforts to standardize data collection. Different data characteristics, species traits, or geographic contexts could yield different relative performance. Testing on additional independent benchmarks would strengthen confidence in these findings.

The finetuning evaluation, while using 30% held-out species, still trained on species from the same regions and environmental contexts. Generalization to truly novel species or biogeographic regions requires further validation.

Computational requirements, though modest by deep learning standards, still include GPU access for finetuning. Pretrained TabPFN without finetuning – which achieves 0.760 median ROC-AUC with ensemble class balancing – remains the most accessible option for practitioners without GPU resources.

While finetuned TabPFN achieves strong ratio calibration (Miller slope = 1.110), absolute calibration remains unachievable for all presence-only models because true prevalence is unknown. Applications requiring absolute occurrence probabilities rather than relative comparisons still need prevalence estimation from independent data sources.

Future Directions

The success of generic TabPFN with domain-specific finetuning suggests even greater potential for PFNs trained specifically for ecological applications. An SDM-PFN could incorporate ecological priors in its training simulations: realistic environmental response curves with physiological limits and optimal ranges (niche theory), dispersal kernels and spatial autocorrelation (spatial processes), observer behavior and accessibility gradients (sampling mechanisms). By learning from simulations with known ground truth, such a model could become inherently robust to sampling bias rather than requiring post-hoc correction.

Recent PFN extensions demonstrate potential for causal inference from observational data[44]. The emerging focus on causal inference in ecology[45,46] provides theoretical frameworks that could guide development of causally-aware ecological PFNs. The simulation-based training paradigm is particularly well-suited to this challenge, as ground truth causal relationships can be specified in training simulations, allowing the model to learn patterns that distinguish causal from merely correlative relationships.

Foundation models also offer capabilities beyond prediction. TabPFN's transformer architecture creates learned representations that could enable transfer learning from data-rich to data-poor species, addressing a persistent challenge in conservation biology where the species most needing distribution models are often those with the fewest occurrence records (e.g. the 'rare species paradox'[47,48]). TabPFN-2.5 introduces a distillation engine that can export predictions as interpretable decision trees or simple neural networks[20], and

TabPFN models generally support SHAP importance scores, potentially addressing interpretability concerns for conservation applications where stakeholder communication requires explainable predictions.

Foundation Models in Ecology: An Emerging Landscape

TabPFN is not the only foundation model approach being developed for species distribution modeling. The growing interest in applying modern machine learning to ecological problems[49] has spawned several foundation model architectures designed specifically for biodiversity applications. NicheFlow uses flow-based generative models to generate species distributions in both environmental and geographic space, with species embeddings that capture meaningful ecological structure across taxa[50]. The iNaturalist Geomodel employs spatial implicit neural representations to predict species occurrence at global scale, trained on millions of citizen science observations[51]. BioAnalyst takes a multimodal approach, integrating species occurrence records, remote sensing data, and climate variables through transformer architectures for spatiotemporal biodiversity forecasting across Europe[52].

These approaches share a common strategy: learning general patterns from large, diverse datasets that can then be applied to specific prediction tasks without task-specific retraining. On the other hand, they differ fundamentally in architecture and in what ecological structures they encode. NicheFlow explicitly models the environmental niche as a probability distribution, enabling generation of hypothetical niches and ancestral niche reconstruction. The iNaturalist Geomodel learns implicit spatial representations that capture geographic range structure. BioAnalyst integrates multiple data modalities to capture spatiotemporal dynamics.

None of these models, however, employ Prior-Data Fitted Networks. This distinction matters for two reasons. First, PFNs are unique in their explicit Bayesian framing – they learn to approximate posterior distributions over model parameters, providing principled uncertainty quantification that other architectures achieve only through additional mechanisms like ensembling or variational inference. Second, and perhaps more consequentially for ecology, PFNs offer a natural path toward addressing sampling bias through their simulation-based training paradigm. By generating synthetic training data with known sampling mechanisms, a PFN could learn to recognize and correct for the kinds of spatial and taxonomic biases that plague presence-only occurrence data.

The diversity of foundation model approaches being developed for SDM reflects both the field's importance and its difficulty. That researchers are exploring normalizing flows, implicit neural representations, multimodal transformers, and PFNs speaks to the recognition that no single architecture may be optimal for all aspects of the SDM problem. Our results with TabPFN demonstrate that general-purpose foundation models can achieve SOTA SDM performance when appropriately adapted. Whether purpose-built ecological foundation models will exceed this performance – and whether they can address fundamental challenges like sampling bias and causal inference that have long constrained SDM methodology – remains an open and exciting question for the field.

Transfer Learning as the Distinguishing Capability

The holdout-to-independent transfer results are not incidental – they provide direct evidence that TabPFN functions as a foundation model for ecological data rather than merely another classification algorithm. Transfer learning, the ability to apply knowledge from one context to new and different contexts[53], is what separates foundation models from conventional machine learning. Random Forest, BRT, MaxNet, and GAM start from scratch on each dataset; they carry nothing forward.

Our results show that this distinction has measurable consequences. The finetuned model’s resistance to overfitting on biased training data – its smaller performance gap on independent test data, its near-ideal ratio calibration – arises because it does not treat each dataset as an isolated problem. For ecology more broadly, this finding matters because ecological datasets are characteristically small, biased, and expensive to collect[54]. Methods that treat each dataset independently are limited by the data they can see. Foundation models that accumulate knowledge across datasets offer a path to better generalization from limited data. The practical barrier is lower than many ecologists may assume: TabPFN runs from R through reticulate, requires no deep learning expertise, and inference takes seconds.

The most promising path forward may be to combine two complementary strategies. First, collecting more bias-independent distribution data – through structured surveys, systematic monitoring programs, and citizen science initiatives designed to minimize spatial sampling bias – would provide higher-quality training material. Second, finetuning foundation models on the widest possible collection of species distribution data, spanning diverse taxa, geographic contexts, and sampling protocols, could teach these models to separate genuine species–environment signals from the artifacts of opportunistic observation. Our results with two-step finetuning on just 226 species already demonstrate meaningful improvement; scaling to thousands of species across global datasets could yield models that handle sampling bias as a natural consequence of their training rather than requiring explicit correction.

Conclusions

TabPFN achieves SOTA species distribution modeling performance when adapted through ensemble class balancing and domain-specific finetuning. If even for the SOTA discrimination alone, TabPFN warrants attention from the SDM community. Foundation models have transformed language processing and computer vision; these results demonstrate they can similarly advance ecological modeling. TabPFN is not the endpoint but a proof-of-concept: if a general-purpose model achieves SOTA SDM performance through adaptation, purpose-built ecological foundation models could address long-standing challenges in biodiversity science and conservation.

References

1. Sánchez-Mercado, A.Y., and Ferrer-Paris, J.R. (2010). Mapping Species Distributions: Spatial Inference and Prediction by Janet Franklin (2009), xviii + 320 pp., Cambridge University Press, Cambridge, UK. ISBN 9780521876353 (hbk), GBP 70.00; 9780521700023 (pbk), GBP 35.00. Oryx *44*, 615–615.

2. Elith, J., and Leathwick, J.R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annu. Rev. Ecol. Evol. Syst. *40*, 677–697.

3. Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., and Blair, M.E. (2017). Opening the black box: an open‐source release of Maxent. Ecography (Cop.) *40*, 887–893.

4. Phillips, S.J., Anderson, R.P., and Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. Ecol. Modell. *190*, 231–259.

5. Elith, J., Leathwick, J.R., and Hastie, T. (2008). A working guide to boosted regression trees. J. Anim. Ecol. *77*, 802–813.

6. Cutler, D.R., Edwards, T.C., Jr, Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., and Lawler, J.J. (2007). Random forests for classification in ecology. Ecology *88*, 2783–2792.

7. Guisan, A., Edwards, T.C., Jr, and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Modell. *157*, 89–100.

8. Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J.J., and Elith, J. (2022). Predictive performance of presence‐only species distribution models: a benchmark study with reproducible code. Ecol. Monogr. *92*. https://doi.org/10.1002/ecm.1486.

9. Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. (2018). A deep learning approach to species distribution modelling. In Multimedia Tools and Applications for Environmental & Biodiversity Informatics (Springer International Publishing), pp. 169–199.

10. Kellenberger, B., Winner, K., and Jetz, W. (2026). The performance and potential of deep learning for predicting species distributions. Glob. Ecol. Biogeogr. *35*. https://doi.org/10.1111/geb.70184.

11. Pichler, M., and Hartig, F. (2023). Machine learning and deep learning–A review for ecologists. Methods Ecol. Evol. *14*, 994–1016.

12. Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., and Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLoS Comput. Biol. *17*, e1008856.

13. Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Neural Inf Process Syst *35*, 507–520.

14. Shwartz-Ziv, R., and Armon, A. (2022). Tabular data: Deep learning is not all you need. Inf. Fusion *81*, 84–90.

15. Borisov, V., Leemann, T., Sebler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2024). Deep neural networks and tabular data: A survey. IEEE Trans. Neural Netw. Learn. Syst. *35*, 7499–7519.

16. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2108.07258.

17. Müller, S., Hollmann, N., Arango, S.P., Grabocka, J., and Hutter, F. (2021). Transformers can do Bayesian inference. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2112.10510.

18. Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2022). TabPFN: A Transformer that solves small tabular classification problems in a second. arXiv [cs.LG].

19. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. Nature *637*, 319–326.

20. Grinsztajn, L., Flöge, K., Key, O., Birkel, F., Jund, P., Roof, B., Jäger, B., Safaric, D., Alessi, S., Hayler, A., et al. (2025). TabPFN-2.5: Advancing the state of the art in tabular foundation models. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2511.08667.

21. Ng, K., Fong, E., Frazier, D.T., Knoblauch, J., and Wei, S. (2025). TabMGP: Martingale Posterior with TabPFN. arXiv [stat.ME]. https://doi.org/10.48550/arXiv.2510.25154.

22. Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ford, A., Guisan, A., Hijmans, R.J., Huettmann, F., Lohmann, L., et al. (2020). Presence-only and presence-absence data for comparing species distribution modeling methods. Biodivers. Inf. *15*, 69–80.

23. Hijmans, R.J. (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology *93*, 679–688.

24. Valavi, R., Elith, J., Lahoz-Monfort, J.J., and Guillera-Arroita, G. (2018). blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. bioRxiv. https://doi.org/10.1101/357798.

25. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography (Cop.) *40*, 913–929.

26. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 5–32.

27. Gomez-Rubio, V. (2018). Generalized additive models: An introduction with R. J. Stat. Softw.

28. Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv [stat.ML].

29. Sofaer, H.R., Hoeting, J.A., and Jarnevich, C.S. (2019). The area under the precision‐recall curve as a performance metric for rare binary events. Methods Ecol. Evol. *10*, 565–577.

30. Miller, M.E., Hui, S.L., and Tierney, W.M. (1991). Validation techniques for logistic regression models. Stat. Med. *10*, 1213–1226.

31. Phillips, S.J., and Elith, J. (2010). POC plots: calibrating species distribution models with presence-only data. Ecology *91*, 2476–2484.

32. Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). Density ratio estimation in machine learning.

33. Fithian, W., and Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. Ann. Appl. Stat. *7*, 1917–1939.

34. Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., and Yates, C.J. (2011). A statistical explanation of MaxEnt for ecologists: Statistical explanation of MaxEnt. Divers. Distrib. *17*, 43–57.

35. Gutmann, M.U., and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. AISTATS *9*, 297–304.

36. Liu, C., Newell, G., and White, M. (2019). The effect of sample size on the accuracy of species distribution models: considering both presences and pseudo‐absences or background sites. Ecography (Cop.) *42*, 535–548.

37. Rousseau, J.S., and Betts, M.G. (2022). Factors influencing transferability in species distribution models. Ecography (Cop.) *2022*. https://doi.org/10.1111/ecog.06060.

38. Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M.J., and Steyerberg, E.W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. J. Clin. Epidemiol. *74*, 167–176.

39. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F.A. (2020). Shortcut learning in deep neural networks. Nat. Mach. Intell. *2*, 665–673.

40. Eremeev, D., Bazhenov, G., Platonov, O., Babenko, A., and Prokhorenkova, L. (2025). Turning tabular foundation models into graph foundation models. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2508.20906.

41. Hoo, S.B., Müller, S., Salinas, D., and Hutter, F. (2026). From tables to time: Extending TabPFN-v2 to time series forecasting. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2501.02945.

42. Zheng, J., Gordon, C., Ji, Y., Saratchandran, H., and Lucey, S. (2025). From tables to signals: Revealing Spectral Adaptivity in TabPFN. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2511.18278.

43. Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. *19*, 181–197.

44. Swelam, O., Purucker, L., Robertson, J., Raum, H., Boedecker, J., and Hutter, F. (2025). Does TabPFN Understand Causal Structures? arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2511.07236.

45. Arif, S., and MacNeil, M.A. (2023). Applying the structural causal model framework for observational causal inference in ecology. Ecol. Monogr. *93*. https://doi.org/10.1002/ecm.1554.

46. Schrodt, F., Beck, M., Estopinan, J., Bowler, D.E., Fontaine, C., Gaüzère, P., Goury, R., Grenié, M., Martins, I.S., Morueta-Holme, N., et al. (2025). Advancing causal inference in ecology: Pathways for biodiversity change detection and attribution. Methods Ecol. Evol. *16*, 2276–2304.

47. Ovaskainen, O., Winter, S., Tikhonov, G., Abrego, N., Anslan, S., deWaard, J.R., deWaard, S.L., Fisher, B.L., Furneaux, B., Hardwick, B., et al. (2025). Common to rare transfer learning (CORAL) enables inference and prediction for a quarter million rare Malagasy arthropods. Nat. Methods *22*, 2074–2082.

48. Lomba, A., Pellissier, L., Randin, C., Vicente, J., Moreira, F., Honrado, J., and Guisan, A. (2010). Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. Biol. Conserv. *143*, 2647–2657.

49. Beery, S., Cole, E., Parker, J., Perona, P., and Winner, K. (2021). Species distribution modeling for machine learning practitioners: A review. In ACM SIGCAS Conference on Computing and Sustainable Societies (ACM), pp. 329–348.

50. Dinnage, R. (2024). NicheFlow: Towards a foundation model for Species Distribution Modelling. bioRxiv, 2024.10.15.618541. https://doi.org/10.1101/2024.10.15.618541.

51. Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S., and Mac Aodha, O. (2023). Spatial Implicit Neural Representations for global-scale species mapping. arXiv [cs.LG], 6320–6342.

52. Trantas, A., Mensio, M., Stasinos, S., Gribincea, S., Khan, T., Podareanu, D., and van der Veen, A. (2025). BioAnalyst: A Foundation Model for biodiversity. arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2507.09080.

53. Hendrycks, D., Lee, K., and Mazeika, M. (2019). Using pre-training can improve model robustness and uncertainty. arXiv [cs.LG]. https://doi.org/10.48550/arXiv.1901.09960.

54. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. Nat. Commun. *13*, 792.