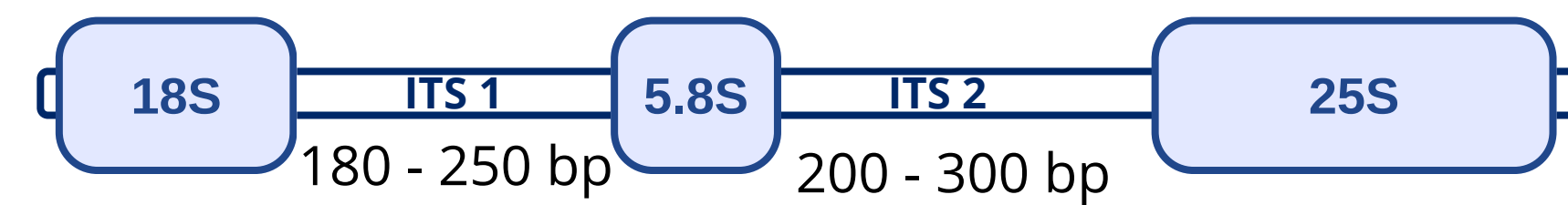


I. Context

- Plant Threat:** 2 out of 5 plant species face extinction [Antonelli et al., 2020].
- Biodiversity Monitoring:** Requires accurate monitoring, harmonized protocols, and quality data [Kolter & Gemeinholzer, 2021].
- Metabarcoding Challenges:** Issues arise from plant genomic complexities like hybridization and polyploidy [Deiner et al., 2017].
- Multimodal Approach:** UMMISCO develops a deep learning model to integrate genomic data and solve taxonomy issues.
- METAPLANTCODE:** Focuses on ITS and trnL-P6 barcodes, part of the Biodiversa+ program [Schema 1].

- eDNA Metadata:** Metadata from eDNA samples will be integrated into the deep learning model.

Nuclear Ribosomal Internal Transcribed Spacer



Chloroplast trnL-P6 Loop (UAA)

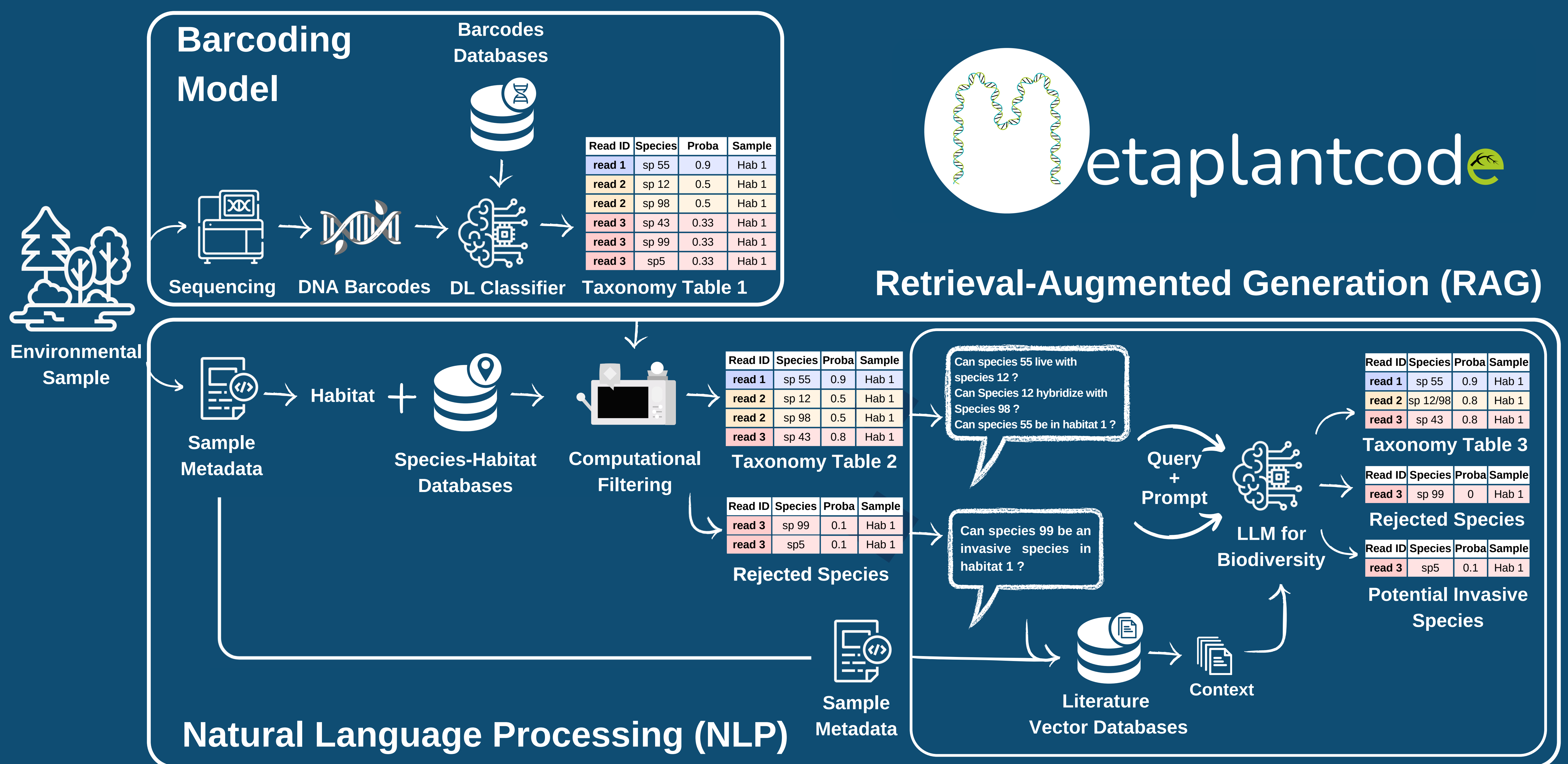


Schema 1: Structural Overview of ITS and trnL Regions in Plant DNA Barcoding.

II. Methods

- Multimodal Approach:** We explored deep learning for barcode classification and we proposed NLP-based metadata integration.
- Dataset:** 156 plant species from the PLANITS database [Banchi et al., 2020], each with 16 barcodes, including reverse complements with no data augmentation.
- Barcoding Model:** We adapted Fluck et al.'s CNN architecture [Flück et al., 2022] for plant barcoding and benchmarked against other models [Zhou et al., 2024; Dalla-Torre et al., 2023; Nguyen et al., 2024; Semlani et al., 2024].

Deep Learning Framework for Multimodal Integration of eDNA Data and Biodiversity Knowledge in Metabarcoding.



III. Results

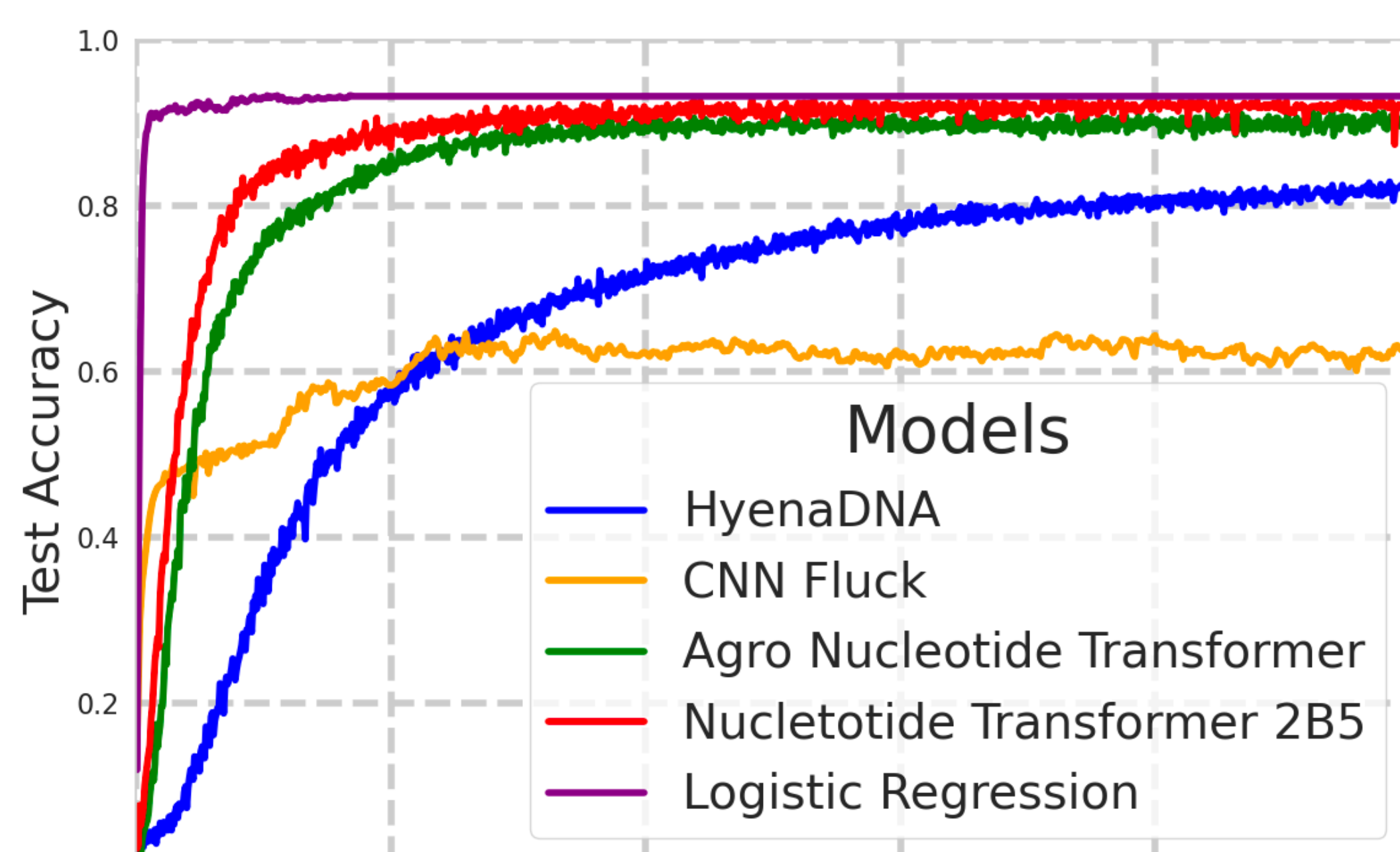


Figure 1: Comparison of Model Test Accuracies on 156 Plant Species

- Fluck's CNN (SOTA)** can be outperformed by genomic foundation models and logistic regression approaches [fig 1].
- Plant-specific foundation models** did not outperform larger, more generalized genomic models [fig 1].
- Classical machine learning (SVM)** performed well on a limited barcoding classification task involving 156 species [table 1].
- Classic bioinformatics tools like BLAST** can be outperformed by newer methods.[table 1].

5. **Our ensemble approach** outperformed the other models, likely due to the distinct error patterns of each individual model [table 1].

Models	Accuracy	F1 Score	Recall	Precision
Logistic Regression	0.938	0.939	0.939	0.949
SVM	0.941	0.941	0.940	0.950
CNN-Fluck	0.632	0.613	0.632	0.651
DNABERT	0.904	0.839	0.840	0.845
DNABERT 2	0.535	0.386	0.398	0.394
DNABERT S	0.825	0.731	0.739	0.733
AgroNucleotide Transformer	0.893	0.816	0.824	0.817
Nucleotide Transformer	0.917	0.868	0.873	0.871
HyenaDNA	0.821	0.718	0.723	0.723
GNN	0.843	0.840	0.843	0.876
Ensemble Method	0.987	0.986	0.987	0.989
BLAST (Bioinformatics)	0.901	0.900	0.901	0.926

Table 1: Benchmark of Basic Bioinformatics, Classical Machine and Deep Learning Models.

IV. Discussion

- We need to **validate our results across 77,000 plant species**, as we anticipate that traditional machine learning methods may struggle to scale effectively.
- Scaling will reveal class **imbalance issues**, necessitating **data augmentation**.
- Benchmark our scaling efforts against **newer, more advanced bioinformatics** pipelines.

References

- Antonelli et al. (2020). State of the World's Plants and Fungi. Royal Botanic Gardens (Kew); Sfumato Foundation.
- Kolter & Gemeinholzer (2021). Plant DNA barcoding necessitates marker-specific efforts. Genome, 64(3), 265-298.
- Deiner et al. (2017). Environmental DNA metabarcoding: Transforming biodiversity surveys. Molecular Ecology, 26(21), 5872-5895.
- Flück et al. (2022). Applying CNNs to accelerate environmental DNA annotation. Scientific Reports, 12(1), 10247.
- Banchi et al. (2020). PLANITS: A curated sequence reference dataset for plant ITS DNA metabarcoding. Database, 2020, baz155.
- Zhou et al. (2023). Dnabert-2: Efficient model for multi-species genome. arXiv:2306.15006.
- Dalla-Torre et al. (2023). The nucleotide transformer: Robust models for human genomics. BioRxiv, 2023-01.
- Nguyen et al. (2024). Hyenadna: Long-range genomic sequence modeling. Advances in Neural Information Processing Systems, 36.
- Semlani et al. (2024). PCN: Deep learning approach to jet tagging. Journal of High Energy Physics, 2024(7), 1-15.

Funding Program

Biodiversita+ is a European Biodiversity Partnership supporting excellent research on biodiversity with an impact for society and policy.