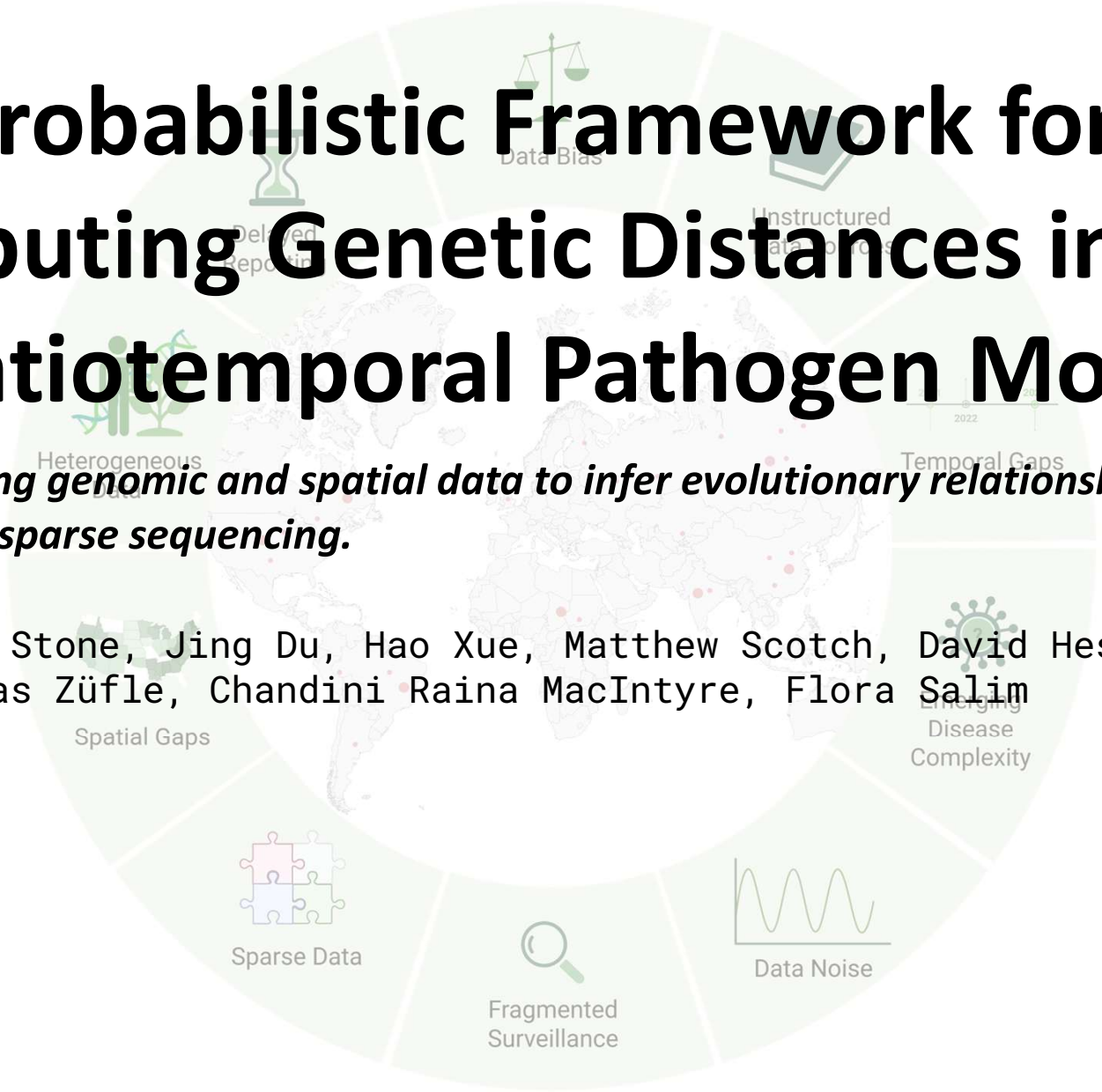
A Probabilistic Framework for Imputing Genetic Distances in Spatiotemporal Pathogen Models

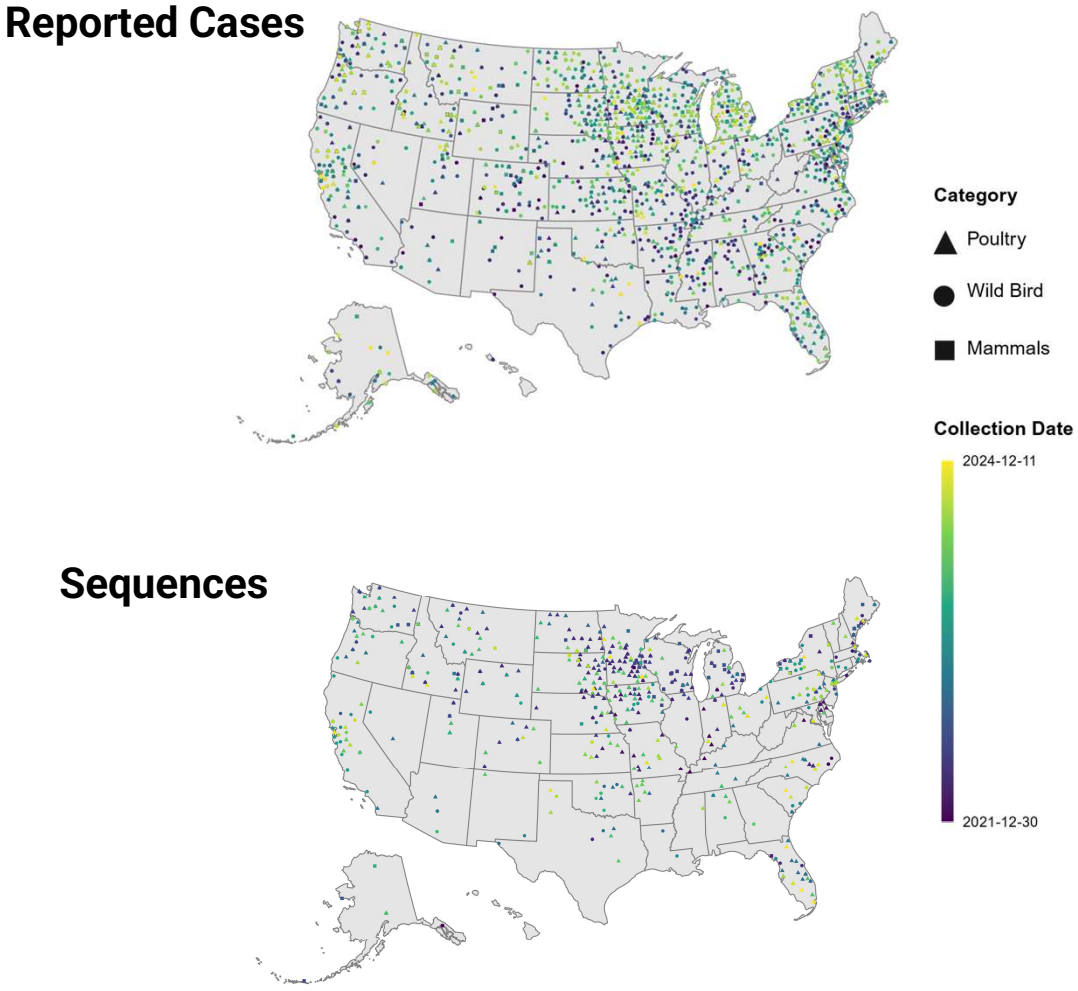
Bridging genomic and spatial data to infer evolutionary relationships under sparse sequencing.

Haley Stone, Jing Du, Hao Xue, Matthew Scotch, David Heslop, Andreas Züfle, Chandini Raina MacIntyre, Flora Salim



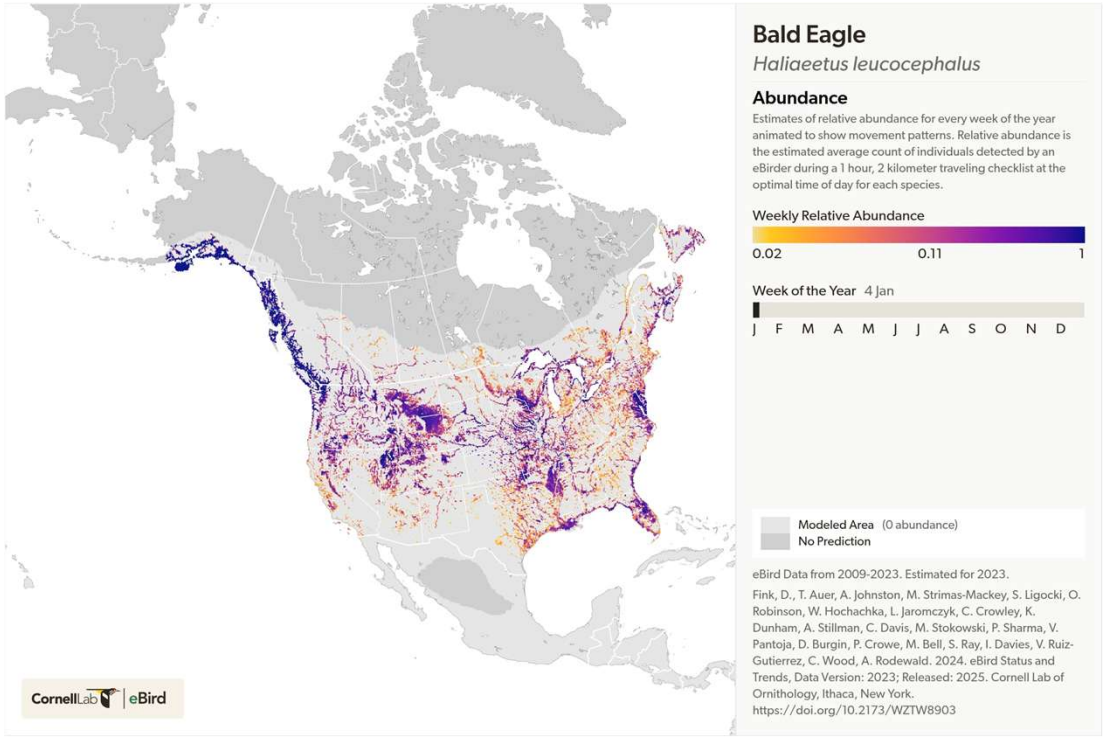
The Problem

- Only a small fraction of detected cases are genetically sequenced
- Without sequences, we lose the ability to measure evolutionary relatedness between detections
- This limits how well spatial models can track viral movement and emergence



Background

- Long-distance migration spreads viruses across continents
- Wildlife–livestock interfaces drive spillover
- Animal surveillance detects risk before human cases appear

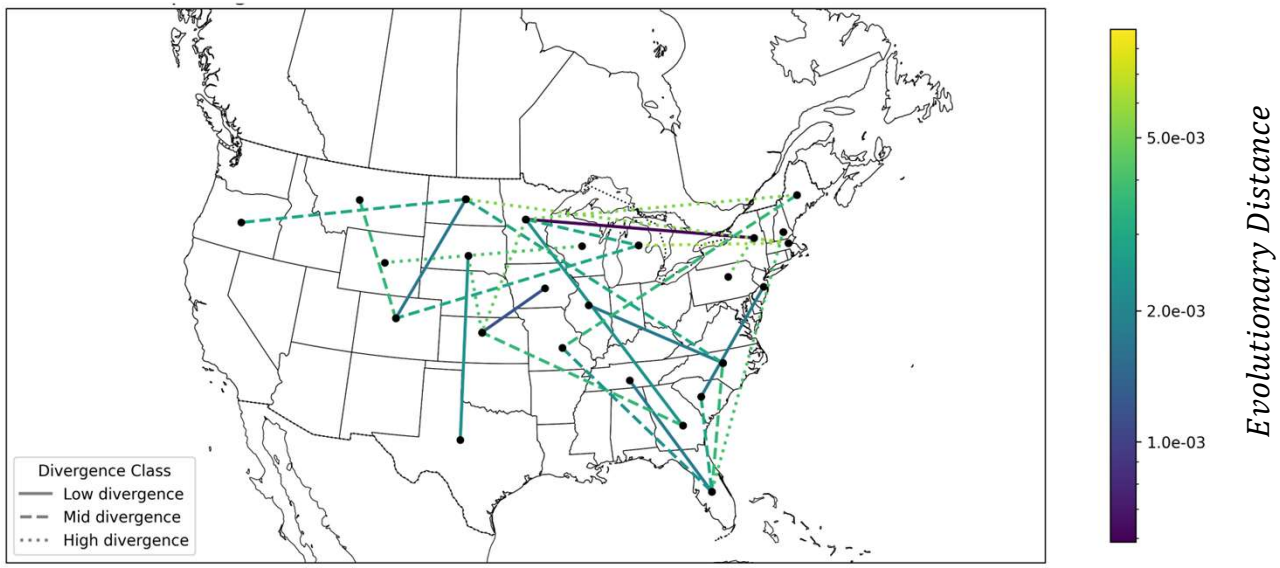


What are evolutionary distances?

- Measure of evolutionary divergence between two genomes
- Counts the number of nucleotide changes, adjusted for multiple substitutions
- Often inferred from a phylogenetic tree as the *path length* between two samples
- Sequence comparisons produce genetic distance matrices that underpin clustering, diffusion, and phylogenetic models

Why Not Just Use Space or Time?

- Without divergence, we lose the evolutionary structure of spread
- Two distant detections can be genetically identical if linked by migration
- Evolution introduces non-Euclidean, asymmetric divergence patterns that simple spatial metrics can’t capture



Problem Formulation

Let $S = \{s_1, \cdots, s_n\}$: sequenced cases
$A = \{a_1, \cdots, a_n\}$: unsequenced cases

For $i \in S \cup A, j \in S$:
metadata vector $x_{ij} \in \mathbb{R}^p$

Observed divergence(only for S x S):
$d_{ij} = K80\ (s_i, s_j)$

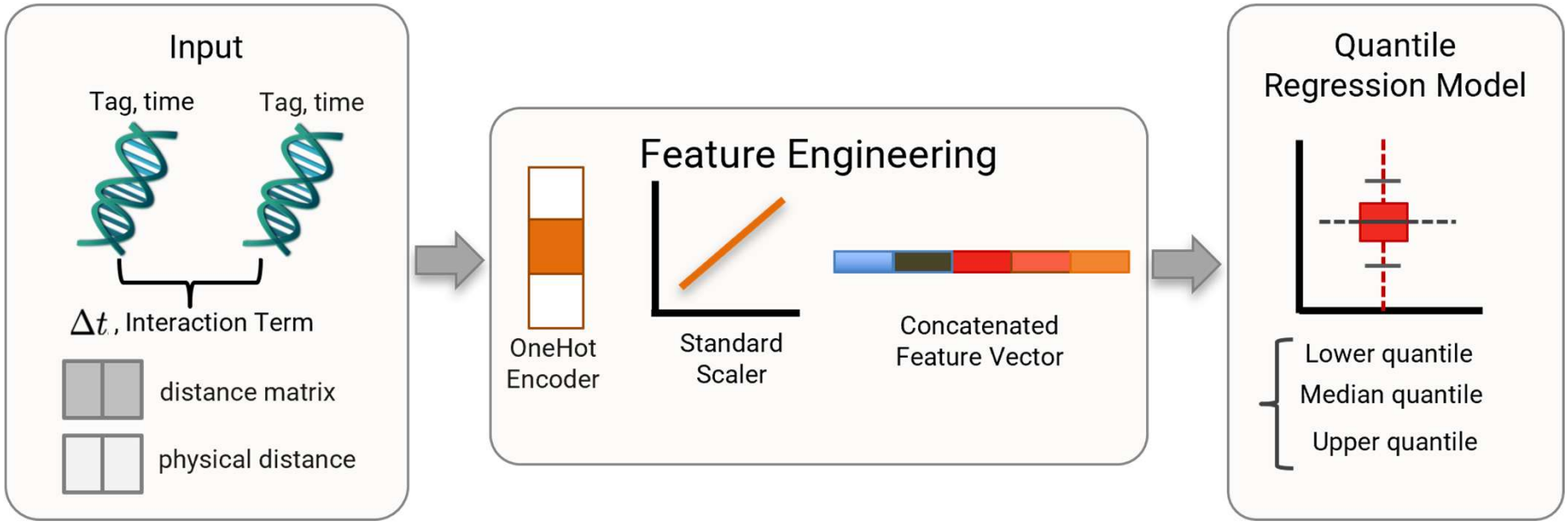
Learn conditional quantiles:
$Q_\tau(d_{ij} \mid x_{ij}), \quad \tau \in \{0.05, 0.5, 0.95\}$

Train on SxS; infer on AxS

UNSW SYDNEY

Metadata-Driven Genetic Distance Imputation

- Learn how metadata explain evolutionary separation between cases
- Train on sequenced pairs to learn the relationship between:
  - Time difference (Δt)
  - Spatial distance (Δgeo)
  - Host taxonomy
  - Year
- Once trained, apply the model to unsequenced detections to infer plausible divergence intervals
- Predicted intervals capture biological uncertainty arising from mutation variability and incomplete sampling



Evolutionary Distance Measure

- We use the Kimura two-parameter model (K80) to quantify genetic divergence
    - Corrects for unobserved mutations and distinguishes transitions vs. transversions
- Provides a biologically meaningful, continuous distance between sequences
- Serves as our ground truth target for learning

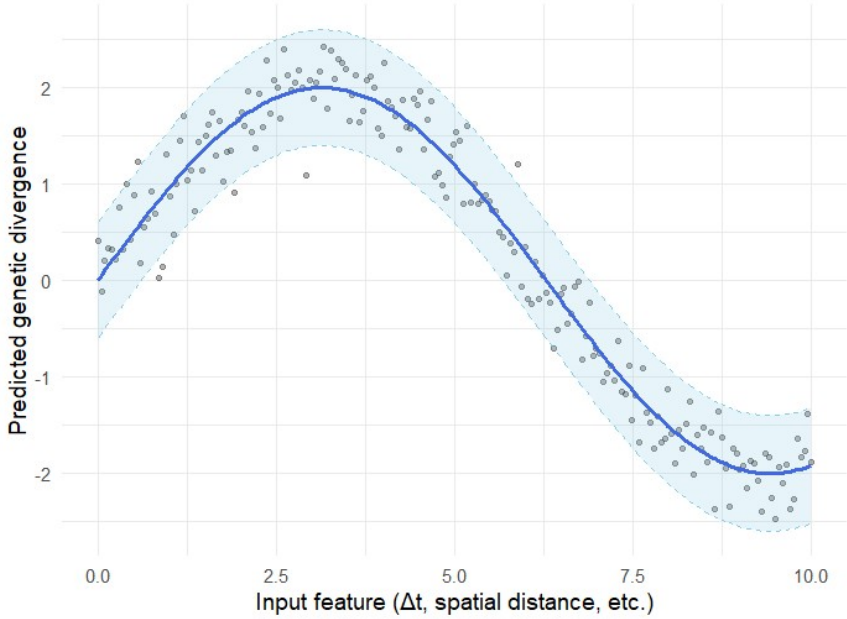
$$d = -\frac{1}{2}\ln(1 - 2P - Q) - \frac{1}{4}\ln(1 - 2Q)$$

UNSW SYDNEY

Quantile Regression Framework

- Three models estimate the 5th, 50th, and 95th percentiles using gradient-boosted trees
- Quantile (pinball) loss allows asymmetric treatment of over- vs under-prediction
- The resulting interval represents expected evolutionary variability

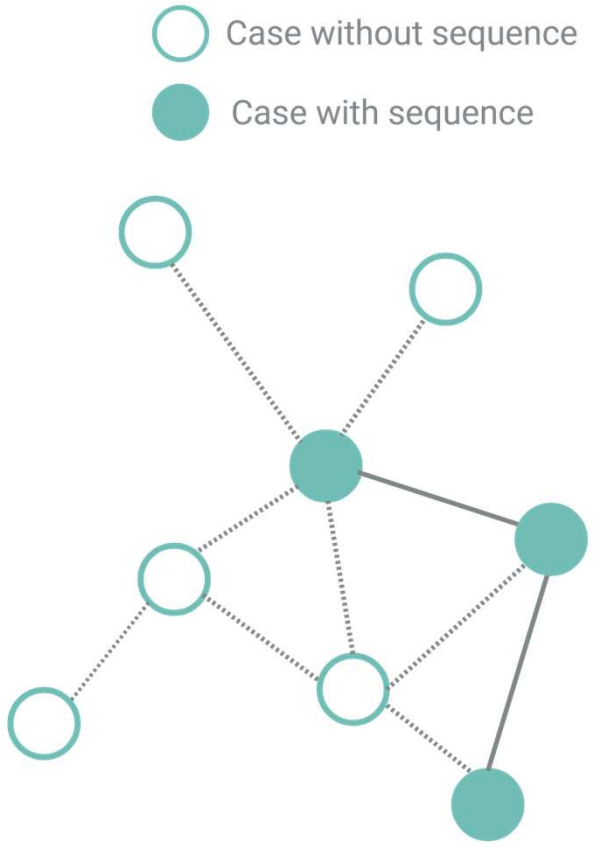
Class	Distance Range	Mean Distance
LOW	[0.000000, 0.003246]	≈0.0020
MID	[0.003246, 0.006632]	≈0.0045
HIGH	[0.006632, 0.492703]	≈0.0193



UNSW SYDNEY

Data & Training Setup

- *AvianUS* Dataset:
  - 2,143 wild bird detections (2021–2024)
  - ~1 M valid sequence pairs used for training
- Cross-validated by region and year to avoid spatial autocorrelation
- Model then applied to all unsequenced detections for nationwide inference

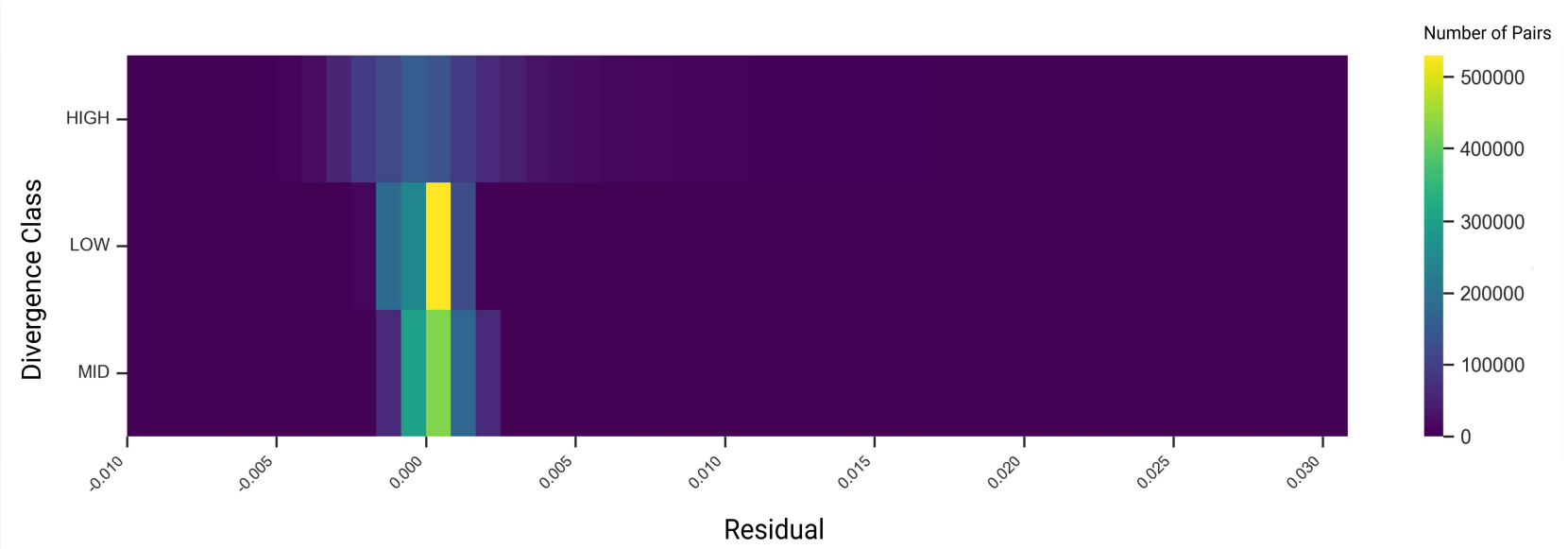


UNSW SYDNEY

Model Performance

- Model reproduces observed divergence patterns across evolutionary scales
- Intervals are sharp and well-calibrated & uncertainty reflects real biological variability
- Strongest correspondence in local and regional transmission contexts

Class	MAE	RMSE	R^2	Coverage (%)
LOW	0.00060	0.00077	0.03	94.95
MID	0.00064	0.00080	0.07	92.63
HIGH	0.01074	0.04352	0.03	90.07
TOTAL	0.00283	0.00542	0.19	92.60

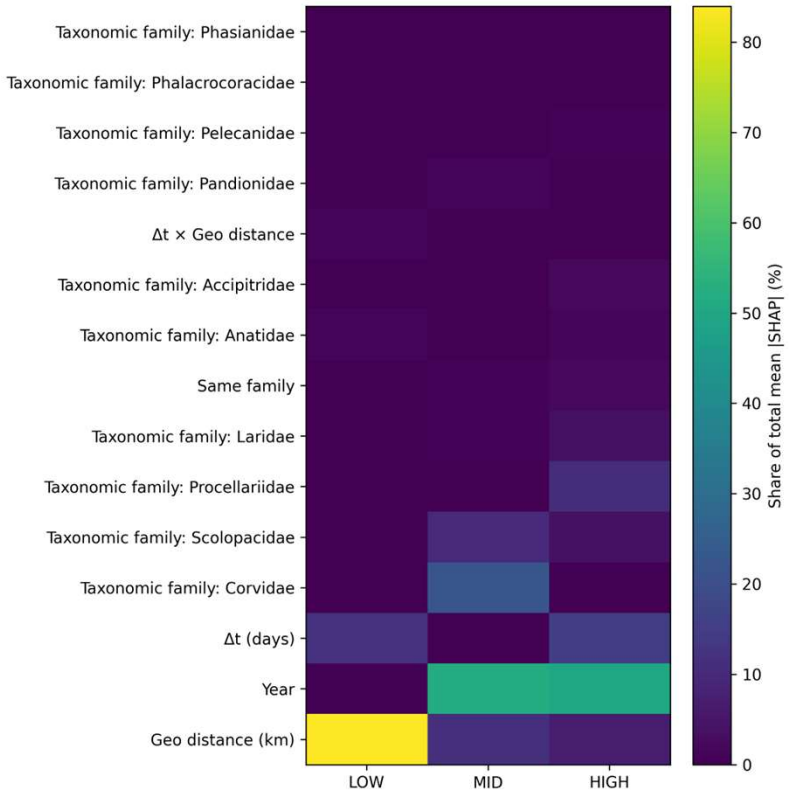


Feature Effects & Interpretability

- *Low divergence:* spatial and temporal separation dominate
- *Mid divergence:* time × host interactions gain influence
- *High divergence:* host taxonomy and year explain most variation
- The model learns biologically interpretable transitions: from within-lineage drift to cross-lineage diversity

Ablation Study

Feature Removed	ΔMAE	ΔInterval Width
Δt	+0.00021	+0.00085
Physical Distance	+0.00013	+0.00049
Same Taxonomic Family	+0.00009	+0.00026
Interaction Term	+0.00015	+0.00037
Year	+0.00007	+0.00031



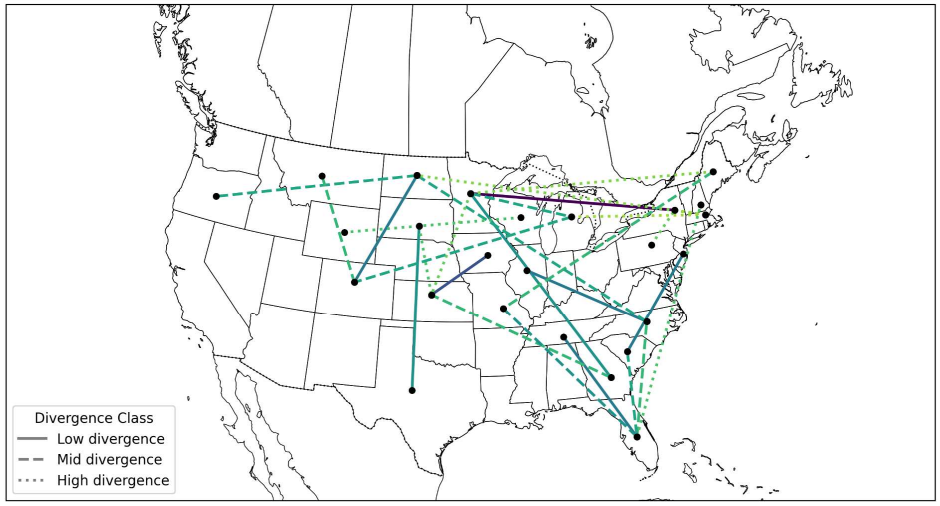
SHAP

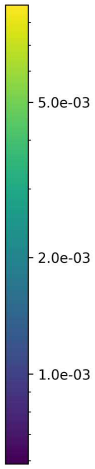
UNSW SYDNEY

Case Study

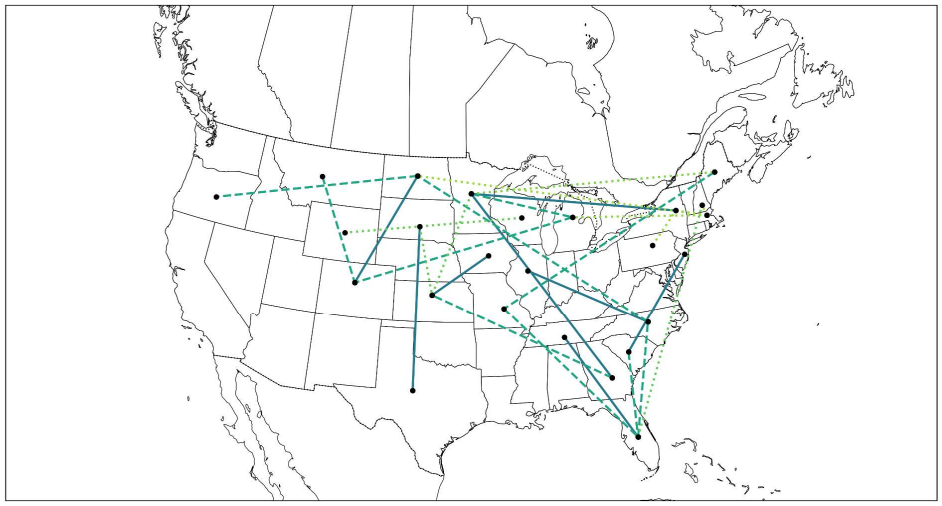
- Observed vs. predicted divergence maps show parallel gradients across regions
- Predicted divergence mirrors observed evolutionary gradients across space and host ecology
- Accurately reflects local transmission, regional mixing, and independent introductions

Observed Evolutionary Distance

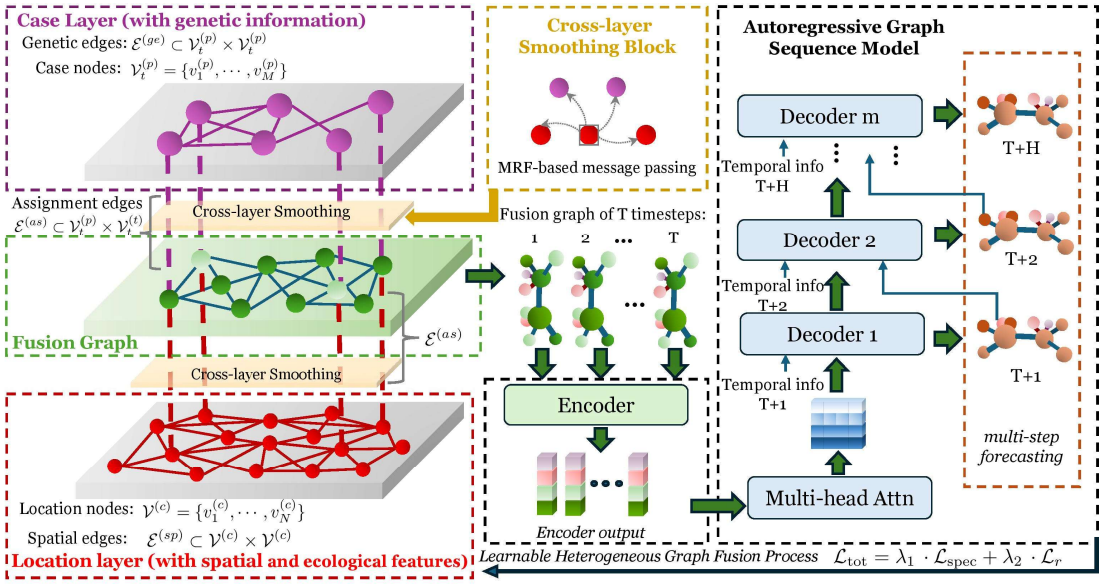




Predicted Evolutionary Distance



Integrating Imputed Distances into Spatiotemporal Graphs



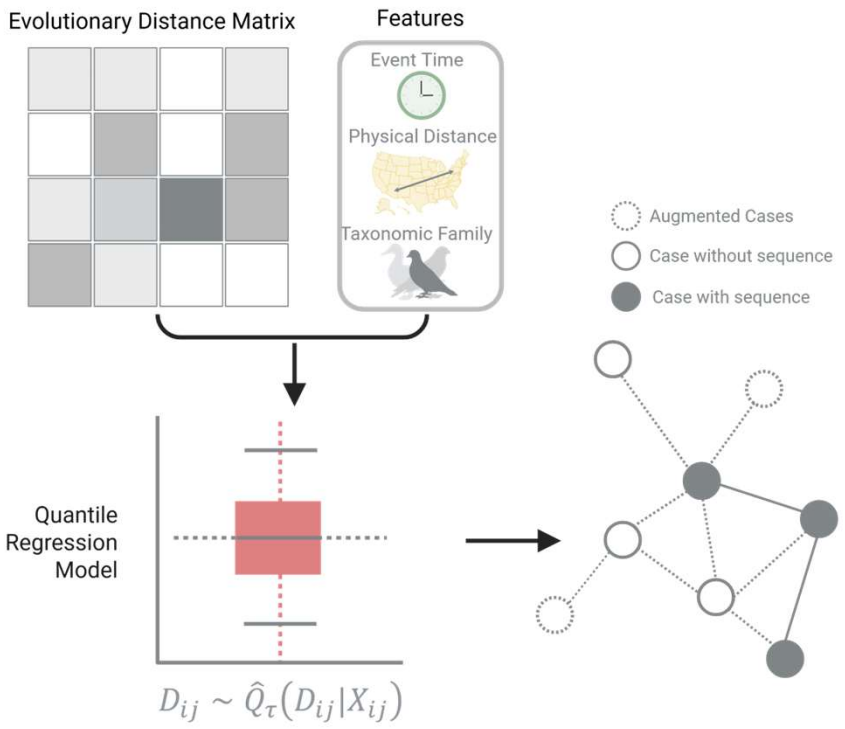
- Each node = detection; edges weighted by observed or imputed divergence
- Supports diffusion modelling and cluster detection under missing data
- Supports our broader BLUE framework for pathogen forecasting

Du, Jing, Haley Stone, Yang Yang, Ashna Desai, Hao Xue, Andreas Züfle, Chandini Raina MacIntyre, and Flora D. Salim. "BLUE: Bi-layer Heterogeneous Graph Fusion Network for Avian Influenza Forecasting." arXiv preprint arXiv:2505.22692 (2025).

UNSW SYDNEY

Summary

- Metadata-driven imputation of genetic distances under missing data
- Quantile regression provides interpretable, uncertainty-aware predictions
- Scales to continental surveillance systems with minimal sequence coverage



Acknowledgements

Funding: *Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO)* & *The United States National Science Foundation* under

- "Collaborative Research: NSF-CSIRO: HCC: Small: Understanding Bias in AI Models for the Pre-diction of Infectious Disease Spread"

Collaborators:

University of New South Wales

- Dr. Jing Du
- Dr. Hao Xue
- Associate Professor David Heslop
- Professor Chandini Raina MacIntyre
- Professor Flora Salim

Arizona State University

- Professor Matthew Scotch

Emory University

- Associate Professor Andreas Züfle

Link to the Git

Link to the Paper