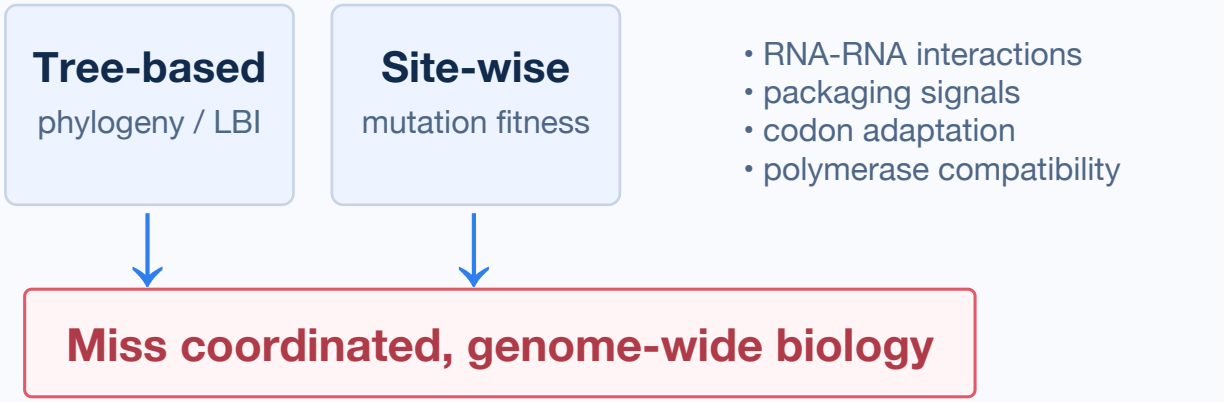
AntigenLM: Structure-aware DNA language modeling for influenza

Yue Pei · Xuebin Chi · Yu Kang

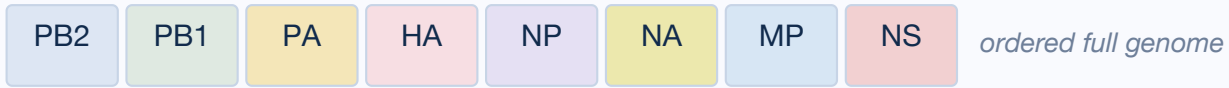
SUMMARY

Why current HA/NA forecasting remains limited

Current predictors still miss coordinated, genome-wide biology.



Influenza A is compact enough for whole-genome DNA modeling (~13k nt)



1 Current predictors remain limited

Tree-based and site-wise predictors often miss coordinated change across HA, NA, and internal segments.

2 Protein-only views miss key signals

Protein-only views miss synonymous, noncoding, RNA-structure, packaging, and codon-level signals.

3 DNA modeling is feasible

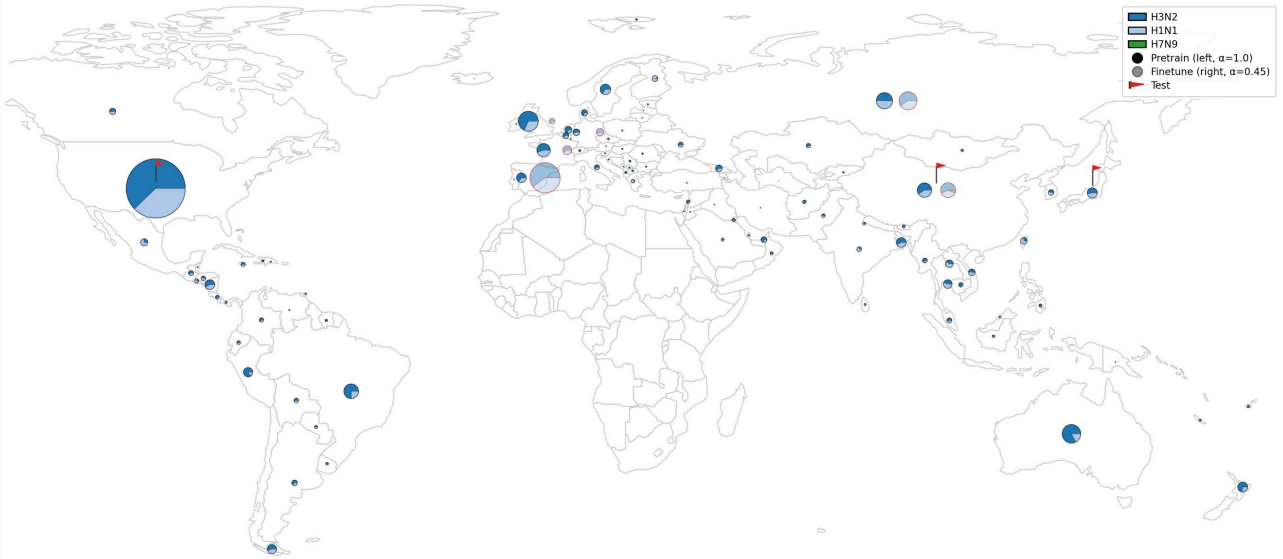
Influenza A is compact enough for whole-genome nucleotide modeling at roughly 13k nt.

Better forecasting needs models that preserve biological structure, not just larger models.

SUMMARY

Why this matters

The bottleneck is not just model size. It is the loss of biological structure..



Data split: pre-Feb-2022 sequences for pretraining/fine-tuning; post-2022 sequences for retrospective testing.

54,512
complete genomes

18
subtypes
~600 million nt

8
influenza A
segments

13k
nt context

1 **Whole-genome coordination**

Packaging signals, polymerase compatibility, and RNA-RNA interactions shape viral evolution beyond HA and NA alone.

2 **Nucleotide-level information matters**

Protein-only models miss synonymous changes, noncoding regulation, and codon-level adaptation signals.

3 **Generic foundation models are not enough**

Large multi-species DNA or protein LMs often preserve local motifs but not influenza-specific global structure.

Genome structure provides a useful inductive bias

1 Whole-genome context

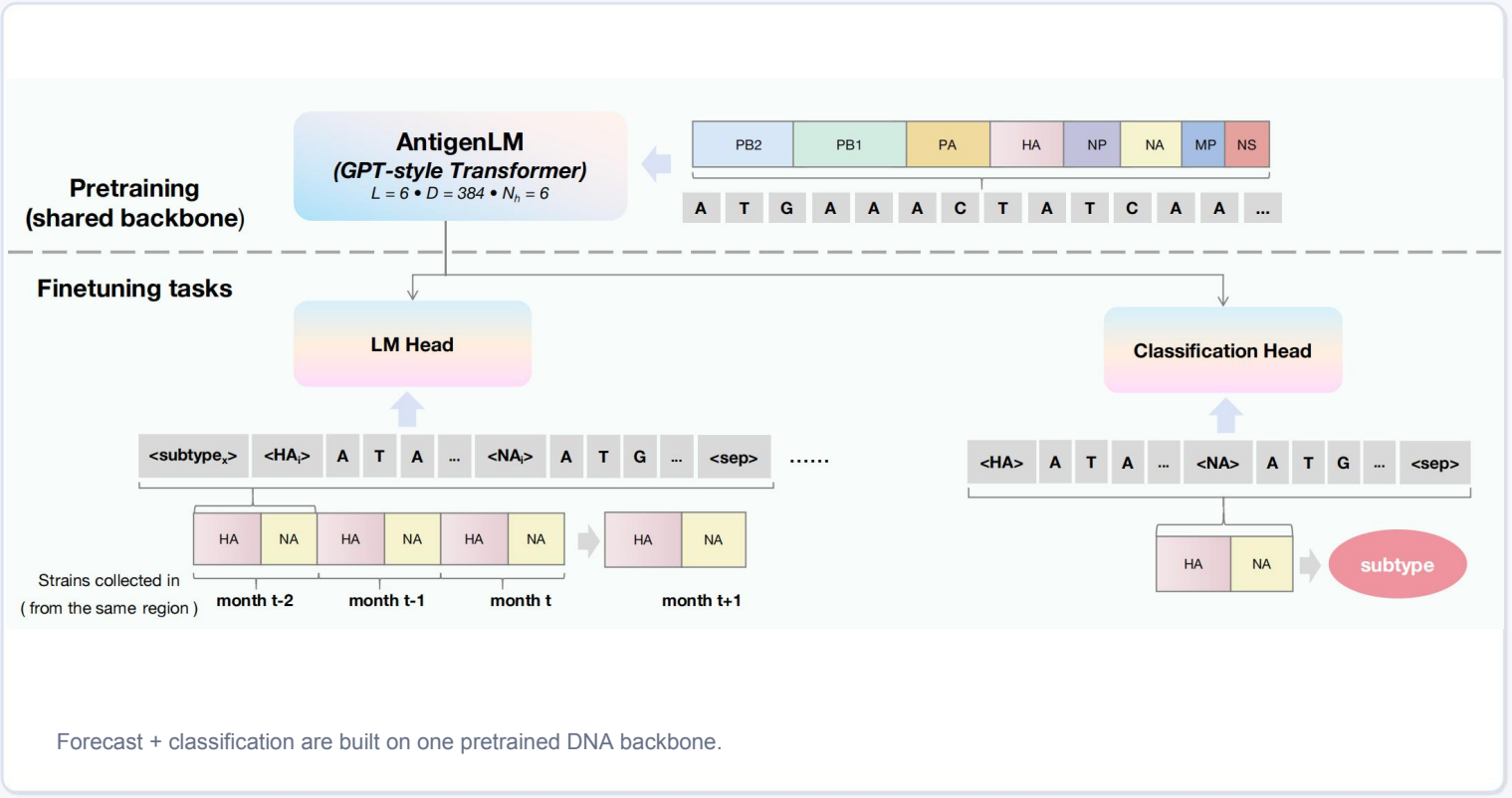
HA, NA and internal segments co-evolve.

2 DNA-level signal

Synonymous and noncoding effects stay visible.

3 One shared backbone

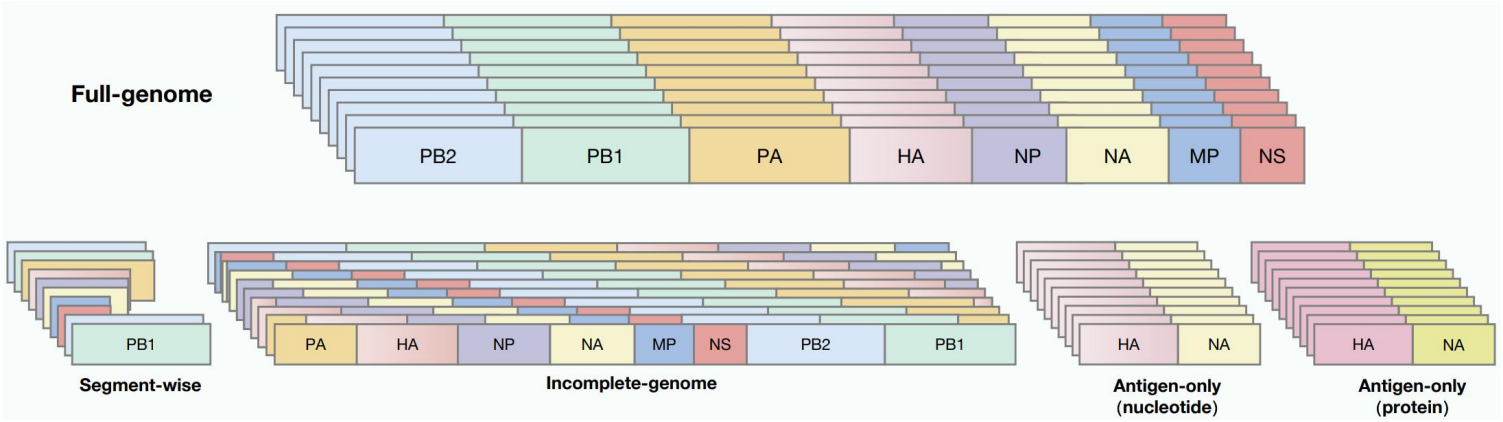
Forecasting and subtype classification share the same pretrained backbone.



Forecast + classification are built on one pretrained DNA backbone.

METHOD

Test the structure hypothesis directly



Ablations isolate the effects of full-genome context, sequence integrity, and antigen-only modeling.

1 Pretrain on ordered full genomes

All 8 segments are kept intact and concatenated in fixed order.

2 Use 3 past HA/NA windows

The model conditions on recent history to generate the future block.

3 Reuse the same backbone

A second head turns the representation into subtype logits.

Full-genome | Segment-wise | Incomplete | Antigen-only

RESULTS

Lower mismatch across forecasting tasks

3-4 AA

HA mismatch
next-month

1-2 AA

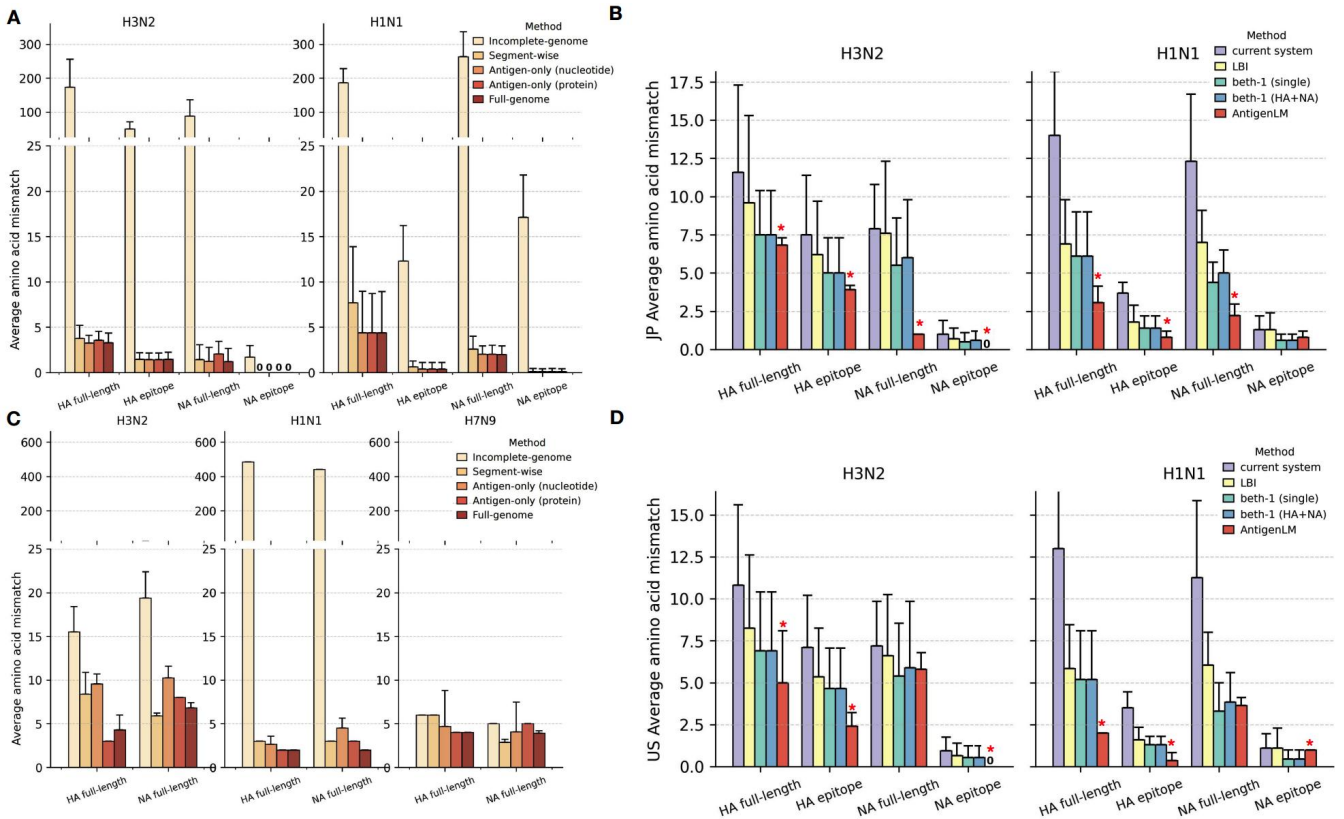
NA mismatch
next-month

>70%

up to lower mismatch
vs current system

PPL 1.26

best among
ablation models



Interpretation

Preserving full-genome structure consistently reduces amino-acid mismatch in both short- and longer-horizon forecasting.

GENERALIZATION

Generalizes beyond the fine-tuning distribution

1 H7N9 transfer

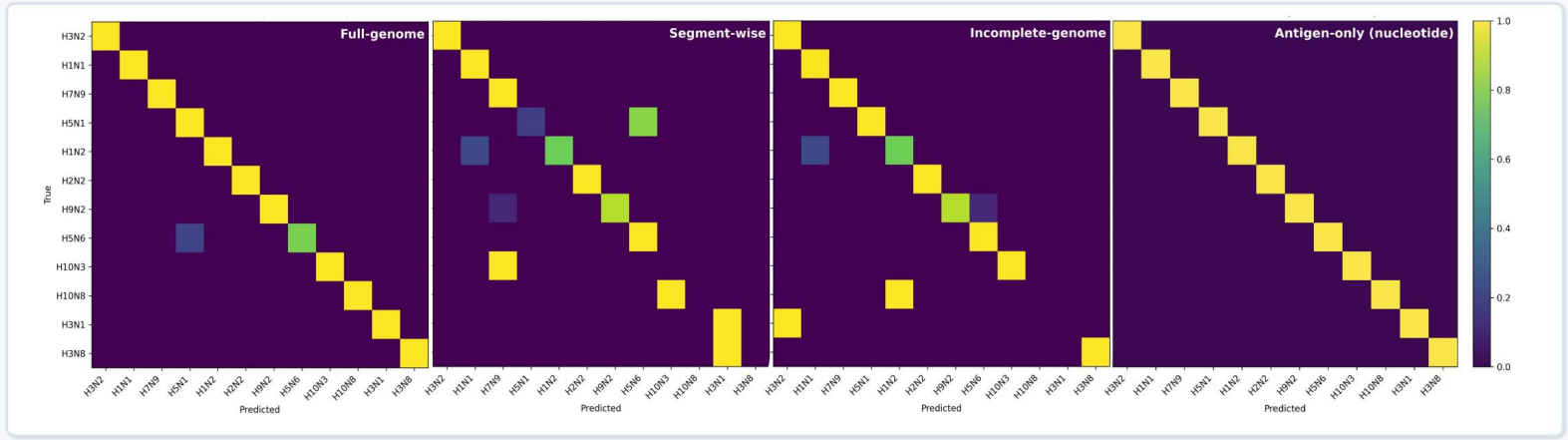
Useful forecasting even with sparse fine-tuning data.

2 Held-out U.S.

HA forecasting gains persist outside the fine-tuning region.

3 99.81% F1

Subtype classification is almost perfect.



4.68%
rare subtype share

48
H7N9 tune set

99.81%
micro-F1

Forecasts are probabilistic and intended to complement, not replace, expert-guided vaccine decisions.

Takeaway: biological structure is not decoration - it is the inductive bias.