

SI Guide

SI Guide for “Standing HA phenotypic breadth shapes H5N1 cross-host potential”

Title	File	Summary
Supplementary Information	Supplementary Information file	Contains 16 supplementary figures, two supplementary table groups and a guide to 21 Supplementary Data files. The display items document phylogenetic, phenotypic, structural, Rosetta, MD and MM/GBSA analyses supporting the manuscript.
Supplementary Data 1: reference_2021_AAseq.fasta	reference_2021_AAseq.fasta	Amino-acid sequence of the 2021 reference HA used to anchor variant comparisons, mutation enumeration and reference-normalized phenotype calculations.
Supplementary Data 2: reference_2021_DNAseq.fasta	reference_2021_DNAseq.fasta	Nucleotide coding sequence of the 2021 reference HA used for alignment, variant mapping and phylogenetic reference coordinates.
Supplementary Data 3: S1.1_master_AA_2195_phenotypes.csv	S1.1_master_AA_2195_phenotypes.csv	One row per unique HA amino-acid variant, with representative accession, counts, sequence-derived phenotypes, structural metrics, stability estimates, receptor-binding estimates and lineage assignments.
Supplementary Data 4: S1.2_master_DNA_13000_phenotypes.csv	S1.2_master_DNA_13000_phenotypes.csv	Per-isolate table for the 13,000-tip HA dataset, including metadata, variant assignment, phenotype values and lineage labels used for tree-level phenotype mapping.
Supplementary Data 5: S1.3_Phenotype_diversity_all_seq.csv	S1.3_Phenotype_diversity_all_seq.csv	Quarterly effective-bin diversity estimates for all sequences, including rarefaction status, raw diversity measures and bootstrap confidence intervals for each phenotype metric.
Supplementary Data 6: S1.4_Phenotype_diversity_lineageA_seq.csv	S1.4_Phenotype_diversity_lineageA_seq.csv	Quarterly effective-bin diversity estimates for Lineage A sequences, including raw and rarefied diversity summaries for each phenotype metric.
Supplementary Data 7: S1.5_Phenotype_diversity_lineageB_seq.csv	S1.5_Phenotype_diversity_lineageB_seq.csv	Quarterly effective-bin diversity estimates for Lineage B sequences, including raw and rarefied diversity summaries for each phenotype metric.
Supplementary Data 8: S1.6_top10_variant_members.csv	S1.6_top10_variant_members.csv	Accession-level membership table for isolates belonging to the ten most frequent HA amino-acid variant classes, including strain, date, variant count, variant label and lineage.
Supplementary Data 9: S1.8_top10_variant_genoflu_counts.csv	S1.8_top10_variant_genoflu_counts.csv	GenoFLU genotype counts for the top ten HA variant classes, including B3.13, D1.1, other assigned genotypes, unassigned records and unknown annotations.
Supplementary Data 10: S2.1_gantt_table.csv	S2.1_gantt_table.csv	Time-binned table underlying variant circulation timelines, with start and end dates, representative accession and host-composition percentages for each top-ten variant.

Title	File	Summary
Supplementary Data 11: S2.2_host_stratified_core_metrics.csv	S2.2_host_stratified_core_metrics.csv	Core phenotype comparisons between avian and domestic-mammal/cattle-associated HA sequences, including medians, interquartile ranges, means, median differences and direction of change.
Supplementary Data 12: S3.1_rbs_area_by_variant_linkage_all10.csv	S3.1_rbs_area_by_variant_linkage_all10.csv	Mean, standard deviation and standard error of MD-derived receptor-binding-site triangle area for each top-ten variant and alpha2,3 or alpha2,6 sialoside linkage.
Supplementary Data 13: S3.2_Af3_rbs_geom_points.csv	S3.2_Af3_rbs_geom_points.csv	AlphaFold3-derived receptor-binding-site geometry points and distance-shift metrics used to summarize pocket opening and structural variation across HA variants.
Supplementary Data 14: S4.1_rbs_all_residues_delta_long.csv	S4.1_rbs_all_residues_delta_long.csv	Per-residue MM/GBSA interaction-energy changes relative to the 2021 reference across receptor-binding-site residues, variants and receptor linkages.
Supplementary Data 15: S4.2_rbs_region_delta_long.csv	S4.2_rbs_region_delta_long.csv	Region-level MM/GBSA interaction-energy changes relative to the 2021 reference for receptor-binding-site loop and helix regions across variants and receptor linkages.
Supplementary Data 16: S4.3_mmgsa_summary_stats_all_cleaned.csv	S4.3_mmgsa_summary_stats_all_cleaned.csv	Mean and standard error of total, gas-phase and solvation MM/GBSA binding-energy components for each top-ten variant and sialoside linkage.
Supplementary Data 17: S4.4_top10_variant_members_final.csv	S4.4_top10_variant_members_final.csv	Final accession-level membership table for top-ten HA variant classes, including strain, date, variant count, variant label and lineage used in downstream summaries.
Supplementary Data 18: GISAID_acknowledgement_table.csv	GISAID_acknowledgement_table.csv	Genome-level acknowledgement table for the 13,000 GISAID records used in the study, including accession IDs, virus names, laboratories, submitters, authors and manuscript metadata.
Supplementary Data 19: rosetta_numbering.csv	rosetta_numbering.csv	Residue-numbering map linking Rosetta pseudo-monomer indices to crystal-structure chain, residue number and residue identity for the HA trimer model.
Supplementary Data 20: rosetta_benchmarking_all_predictions.csv	rosetta_benchmarking_all_predictions.csv	Predicted and experimental delta-delta-G values for benchmark receptor-binding mutations across Rosetta protocols and sialoside linkages.
Supplementary Data 21: rosetta_benchmarking_stats_overall.csv	rosetta_benchmarking_stats_overall.csv	Protocol-level Rosetta benchmarking statistics, including Pearson and Spearman correlations, confidence intervals and RMSE values used to select the final modeling protocol.