Supplementary Information

Standing HA phenotypic breadth shapes H5N1 cross-host potential

M H M Mubassir[1,2], Sachin Subedi[1,2], Tanin Rajamand[1,2], Mohamed Bakheet[2,3,4], Ludy Registre Carmola[3], Sihua Peng[2,3], Rajan Kandel[1], Guppy Stott[1,2], Robert J Woods[5,6,7], S. Mark Tompkins[3,8], Gerardus Josephus Boons[5,6] and Justin Bahl[1,2,3,4,8 *]

[1]Institute of Bioinformatics, University of Georgia, Athens, GA, United States.
[2]Center for Ecology of Infectious Diseases, University of Georgia, Athens, GA, United States.
[3]Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA, United States.
[4]Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA, United States.
[5]Complex Carbohydrate Research Center, University of Georgia, Athens, GA, United States.
[6]Department of Chemistry, University of Georgia, Athens, GA, United States
[7]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA
[8]Center for Vaccines and Immunology, University of Georgia, Athens, GA, United States

*Corresponding author. Email: justin.bahl@uga.edu

Summary of Supplementary Information items

This Supplementary Information file includes 16 supplementary figures, 2 supplementary table groups and a guide to 21 accompanying Supplementary Data files. The figures and tables collectively document the phylogenetic, sequence-based, structural, Rosetta, molecular-dynamics and MM/GBSA analyses supporting the manuscript.

Supplementary Figures

1. Supplementary Fig.1. Phylogenetic and phenotypic landscape of U.S. clade 2.3.4.4b H5 HA variants.
2. Supplementary Fig.2. Biophysical charge metrics across the clade 2.3.4.4b HA phylogeny.
3. Supplementary Fig.3. Hydropathy and aliphatic index metrics across the U.S. clade 2.3.4.4b HA phylogeny.
4. Supplementary Fig.4. Amino-acid composition charge fractions across the U.S. clade 2.3.4.4b HA phylogeny.
5. Supplementary Fig.5. Secondary-structure propensities across the clade 2.3.4.4b HA phylogeny.
6. Supplementary Fig.6. High-confidence AlphaFold3 ensemble and RBS geometry for HA variants.
7. Supplementary Fig.7. AlphaFold3 structural divergence and RBS geometry metrics across the U.S. clade 2.3.4.4b HA phylogeny.
8. Supplementary Fig.8. Benchmarking Rosetta ΔΔG protocols and mapping HA stability and receptor binding on the phylogeny.
9. Supplementary Fig.9. Protein backbone stability across HA-sialoside MD simulations.
10. Supplementary Fig.10. Per-residue RMSF profiles across replicate HA-sialoside MD simulations.
11. Supplementary Fig.11. Lineage-stratified RBS triangle geometry.
12. Supplementary Fig.12. Per-residue HA-sialoside binding energetics across clade 2.3.4.4b variants.
13. Supplementary Fig.13. Variant-specific redistribution of HA-sialoside binding energetics relative to the 2021 reference.
14. Supplementary Fig.14. CA landscapes of MD trajectories for V3 and V10 HA-sialoside complexes.
15. Supplementary Fig.15. PLIP interaction fingerprints for Lineage A (V3) and Lineage B (V10).
16. Supplementary Fig.16. Rosetta cartddg2020 workflow and protocol benchmarking against published wet-lab results.

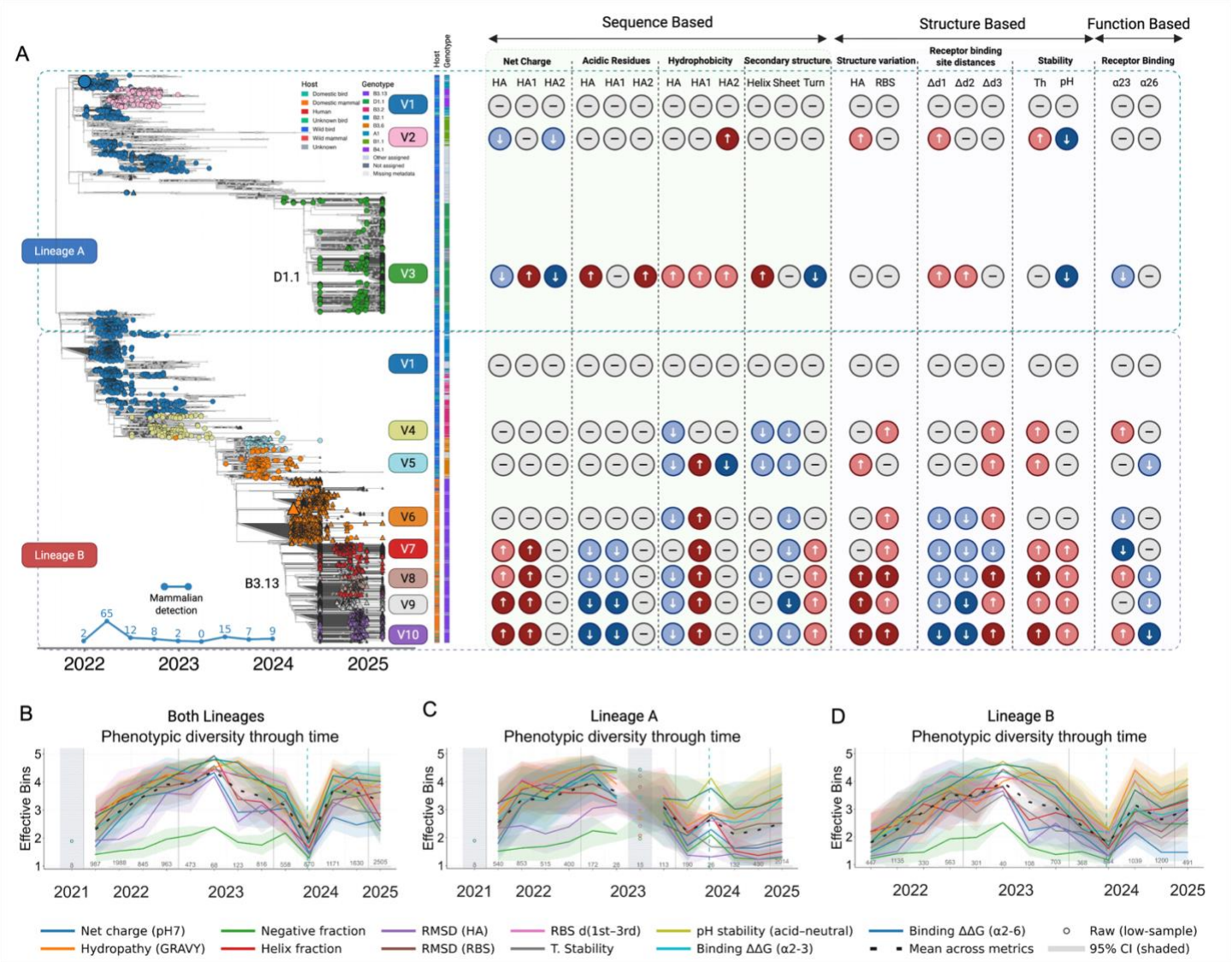
Supplementary Tables

1. Supplementary Table 1. Lineage-stratified phenotype summary. Median and interquartile-range summaries across Lineage A, Lineage B, cattle-associated Lineage B variants and non-cattle Lineage B variants.
2. Supplementary Table 2. Top ten HA variant identities, host composition and phenotype metrics. Includes Supplementary Tables 2a–2f covering variant identity, host composition, sampled species, sequence phenotypes, amino-acid composition, structural metrics, Rosetta estimates and MM/GBSA metrics.

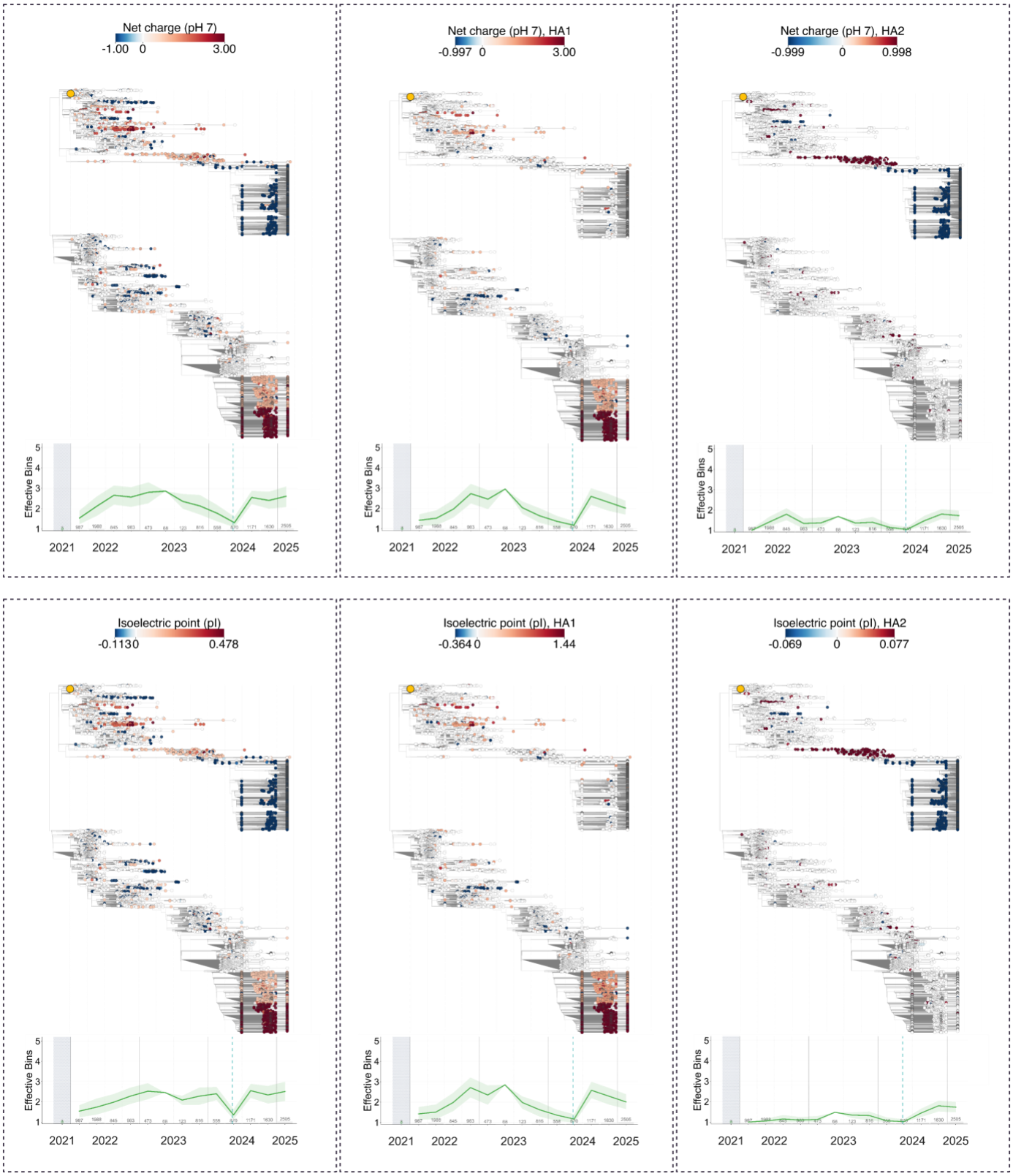
Supplementary Data files

1. reference_2021_AAseq.fasta.
2. reference_2021_DNAseq.fasta.
3. S1.1_master_AA_2195_phenotypes.csv.
4. S1.2_master_DNA_13000_phenotypes.csv.
5. S1.3_Phenotype_diversity_all_seq.csv.
6. S1.4_Phenotype_diversity_lineageA_seq.csv.
7. S1.5_Phenotype_diversity_lineageB_seq.csv.
8. S1.6_top10_variant_members.csv.

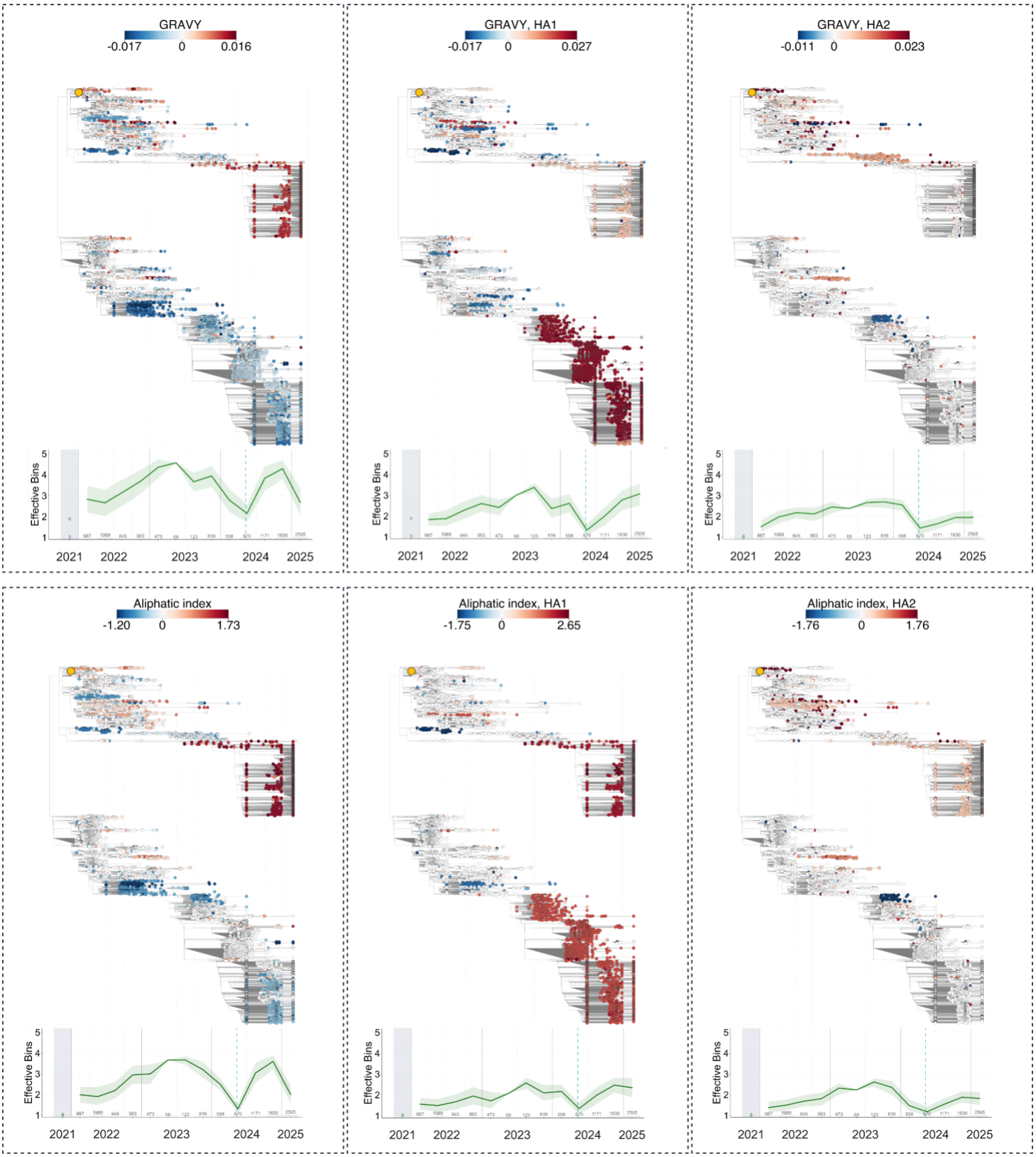
9. S1.8_top10_variant_genoflu_counts.csv.
10. S2.1_gantt_table.csv.
11. S2.2_host_stratified_core_metrics.csv.
12. S3.1_rbs_area_by_variant_linkage_all10.csv.
13. S3.2_Af3_rbs_geom_points.csv.
14. S4.1_rbs_all_residues_delta_long.csv.
15. S4.2_rbs_region_delta_long.csv.
16. S4.3_mmgbsa_summary_stats_all_cleaned.csv.
17. S4.4_top10_variant_members_final.csv.
18. GISAID_acknowledgement_table.csv.
19. rosetta_numbering.csv.
20. rosetta_benchmarking_all_predictions.csv.
21. rosetta_benchmarking_stats_overall.csv.



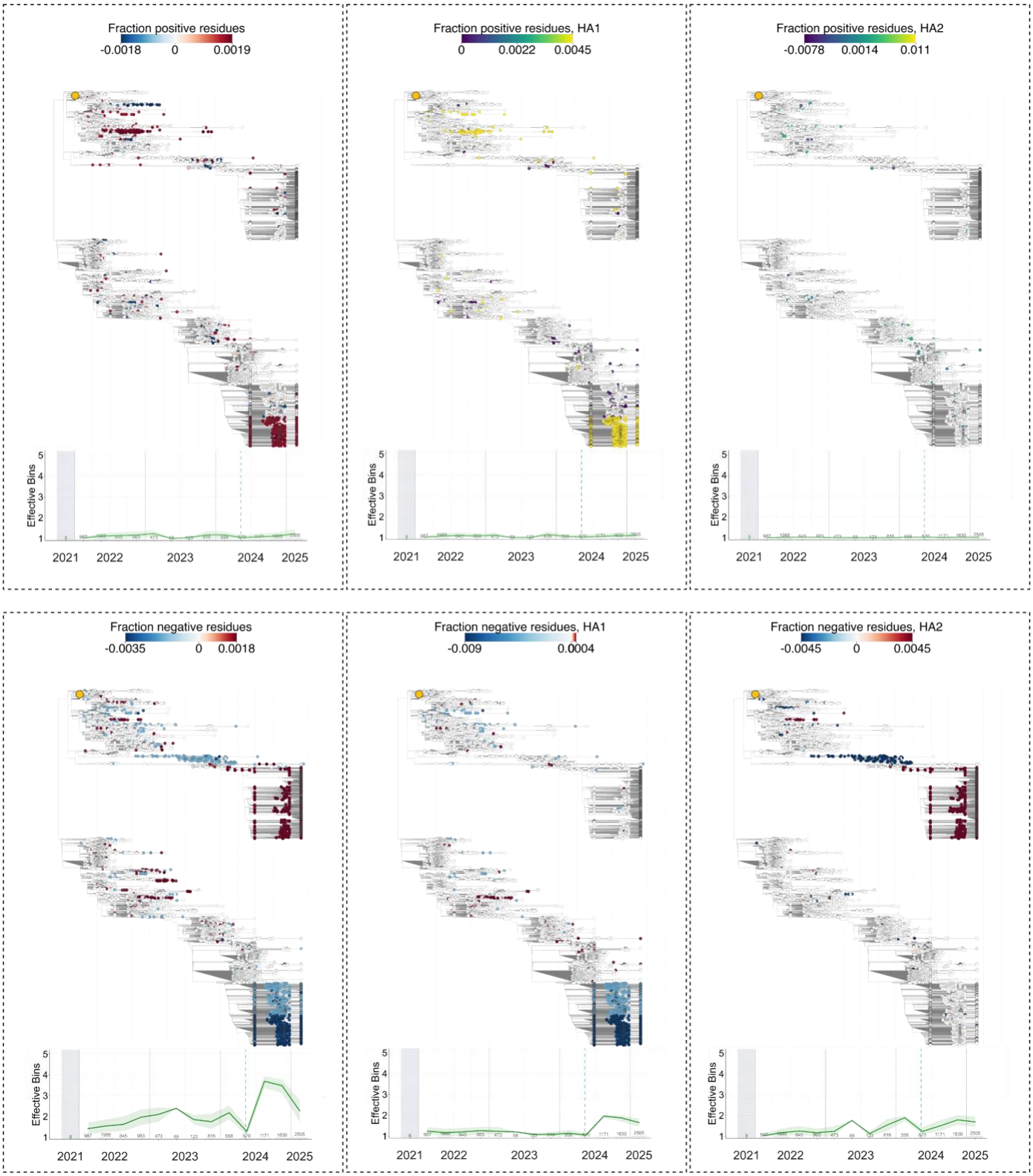
Supplementary Fig.1. Phylogenetic and phenotypic landscape of U.S. clade 2.3.4.4b H5 HA variants. A) Time-scaled HA phylogeny of ~13,000 U.S. H5N1 clade 2.3.4.4b sequences with the ten most frequent HA amino-acid variant classes (V1–V10) highlighted and colored. The large circle marks the 2021 reference virus, and the large triangle marks the early Texas/dairy-cattle reference. Mammalian detections, including domestic mammals, wild mammals, and humans, are shown as triangles; other tips are shown as circles. The two vertical metadata strips at right indicate host class and grouped GenoFLU genotype annotation for each tree tip. Major GenoFLU genotypes are shown separately, and lower-frequency annotations are grouped as “Other assigned” or “Not assigned.” GenoFLU labels are used as contextual whole-genome metadata and were not inferred from HA sequence alone. (B–D) Time-resolved trajectories of phenotypic diversity from 2021–2025, plotted as the number of effective bins. Open circles show raw bin-level estimates, grey shading indicates low-sample/masked bins, and the vertical dashed line marks the cattle-introduction period.



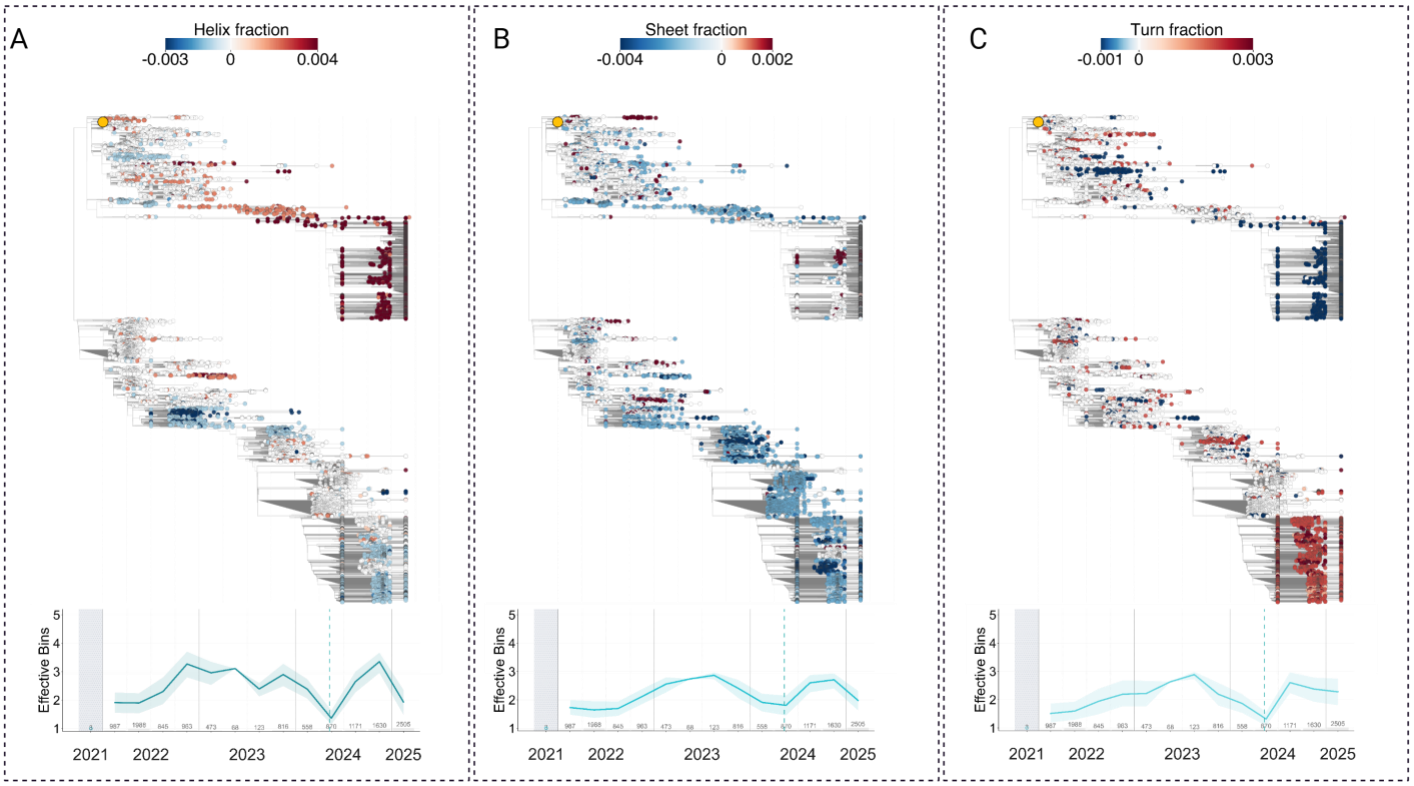
Supplementary Fig.2. Biophysical charge metrics across the clade 2.3.4.4b HA phylogeny. Time-scaled HA trees are colored by sequence-derived net charge at pH 7 (top row) and isoelectric point (pI) (bottom row) for the full HA ectodomain, HA1, and HA2 (left to right). Colors span more negative/lower values (blue) to more positive/higher values (red), and the 2021 wild-bird reference is marked by a yellow circle. Green traces below each tree show the time-resolved effective-bin diversity for the corresponding metric; grey shading indicates low-sample intervals and the dashed vertical line marks the cattle-introduction period.



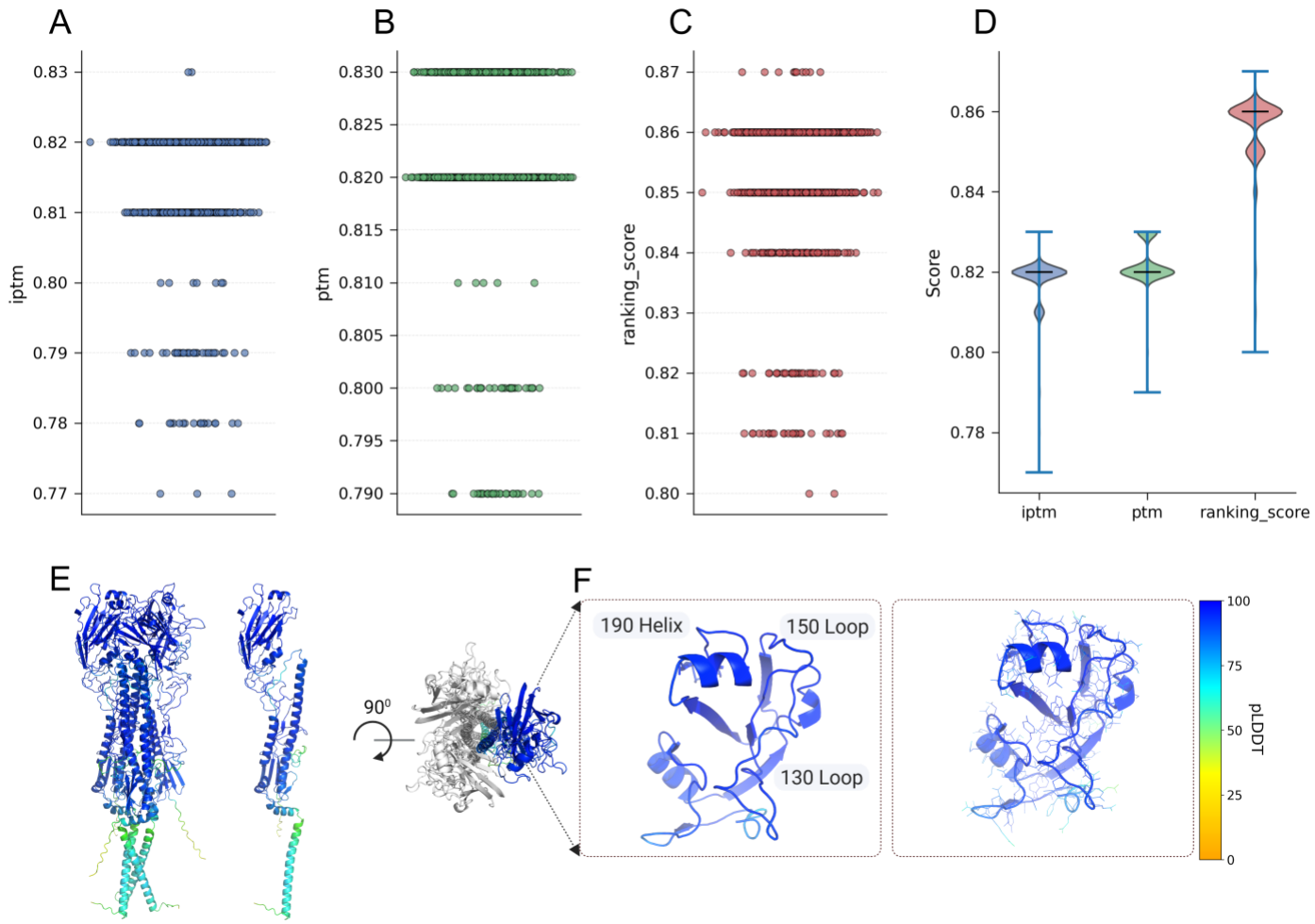
Supplementary Fig.3. Hydropathy and aliphatic index metrics across the U.S. clade 2.3.4.4b HA phylogeny. Time-scaled HA trees are colored by sequence-derived GRAVY (top row) and aliphatic index (bottom row) for the full HA ectodomain, HA1, and HA2 (left to right). Colors span lower to higher values (blue to red), with the 2021 wild-bird reference marked by a yellow circle. Green traces below each tree show the time-resolved effective-bin diversity for each metric; grey shading indicates low-sample intervals and the dashed vertical line marks the cattle-introduction period.



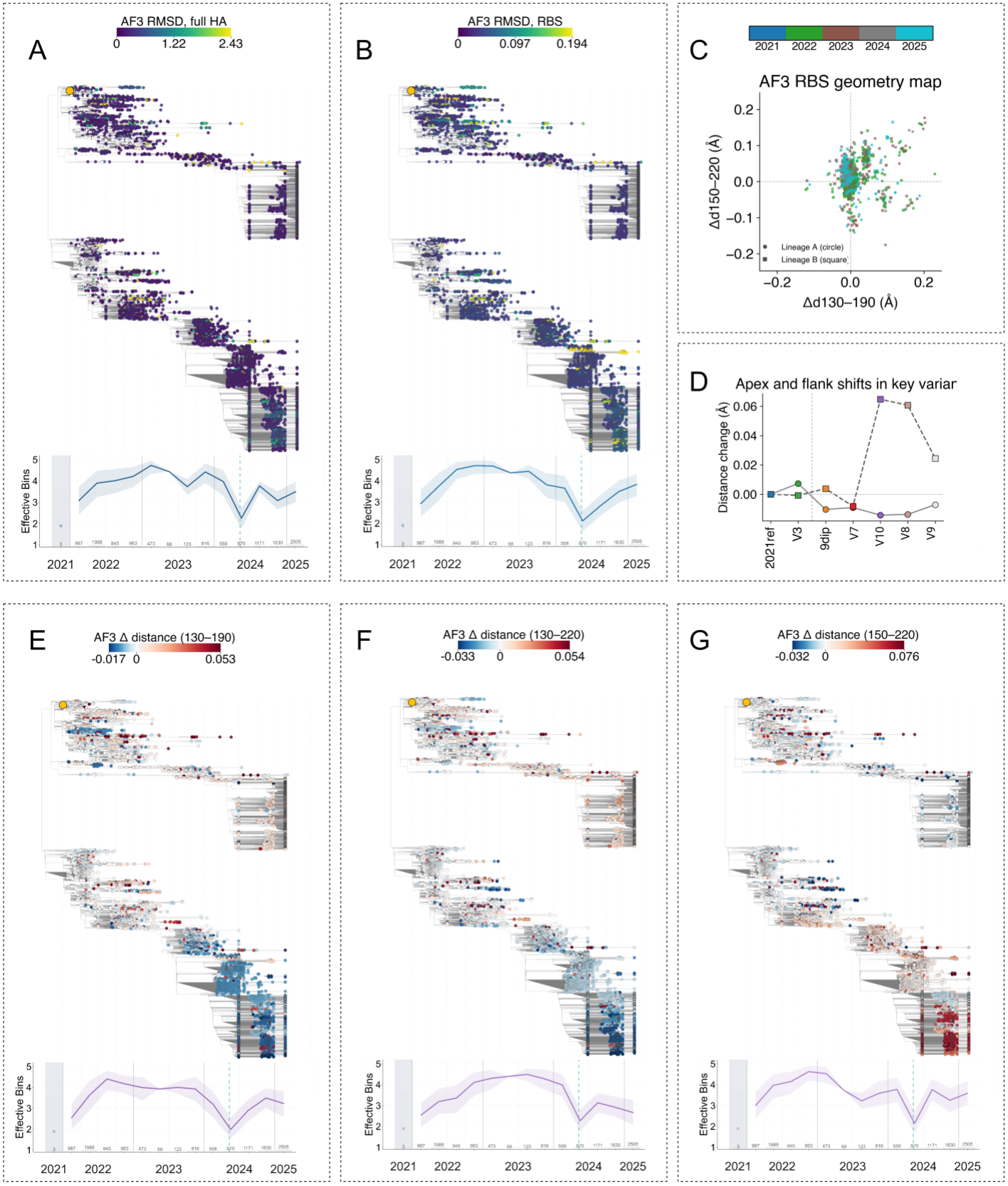
Supplementary Fig.4. Amino-acid composition charge fractions across the U.S. clade 2.3.4.4b HA phylogeny. Time-scaled HA trees are colored by the fraction of positively charged residues (top row) and fraction of negatively charged residues (bottom row), shown for the full HA ectodomain, HA1, and HA2 (left to right) as changes relative to the 2021 wild-bird reference (yellow circle). Color scales indicate lower vs higher values (blue to red or purple to yellow, as shown). Green traces below each tree show the time-resolved effective-bin diversity for the corresponding metric; grey shading denotes low-sample intervals, and the dashed vertical line marks the cattle-introduction period.



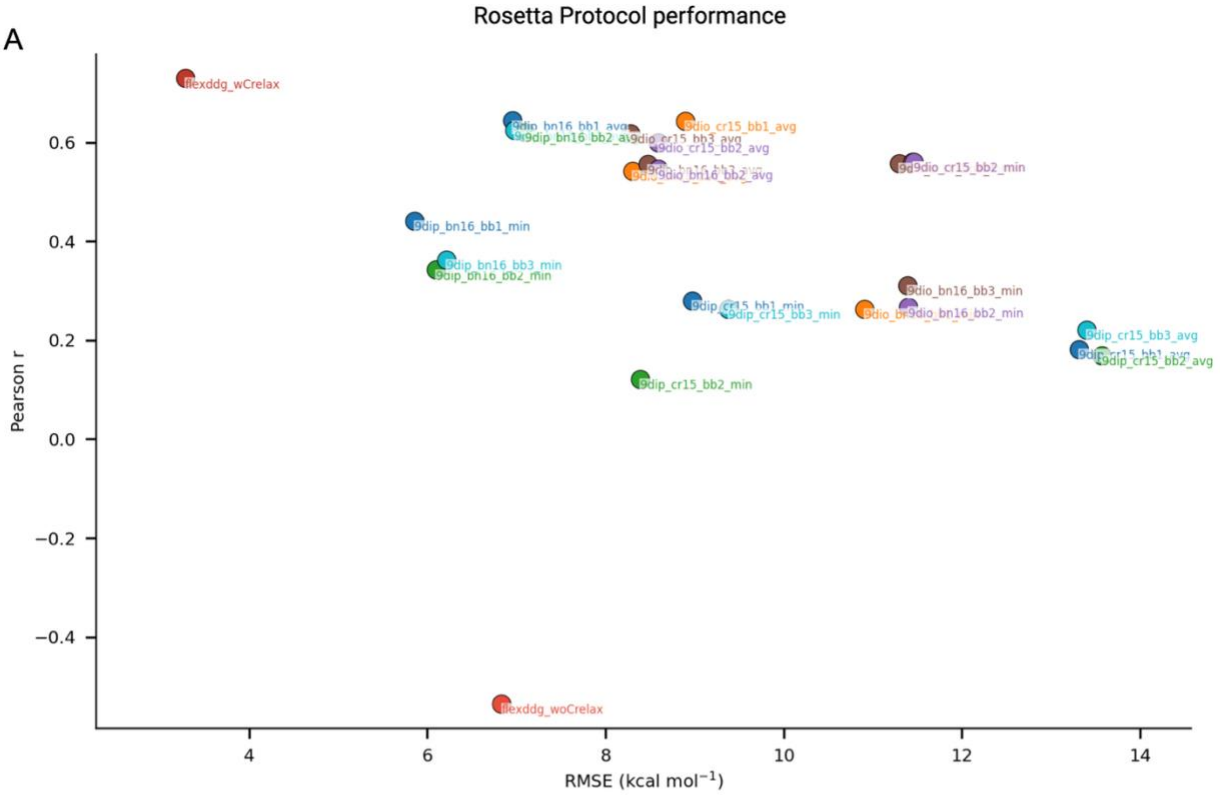
Supplementary Fig.5. Secondary-structure propensities across the clade 2.3.4.4b HA phylogeny. Time-scaled HA trees colored by the fraction of residues predicted to form (A) α-helices, (B) β-sheets, and (C) turns, with colors ranging from lower (blue) to higher (red) fractions. The 2021 wild-bird reference is shown as a large yellow circle. Blue traces below each tree show the time-resolved effective-bin diversity for the corresponding metric; grey shading indicates low-sample intervals, and the dashed vertical line marks the cattle-introduction period.



Supplementary Fig.6. High-confidence AlphaFold3 ensemble and RBS geometry for HA variants. (A–C) ipTM, pTM, and overall ranking_score for AlphaFold3 models of all HA variants. **(D)** Violin plots summarizing the distributions of ipTM, pTM, and ranking_score. **(E)** Representative AF3 trimer for the 2021 reference HA, colored by pLDDT. **(F)** Close-up of the receptor-binding site and associated side chains, colored by pLDDT.



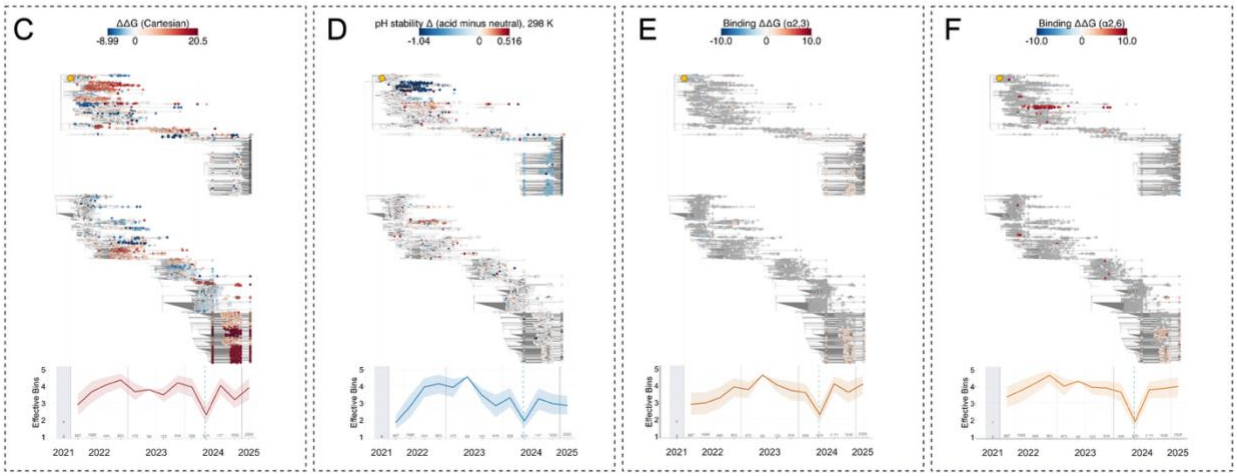
Supplementary Fig.7. AlphaFold3 structural divergence and RBS geometry metrics across the U.S. clade 2.3.4.4b HA phylogeny. (A,B) Time-scaled HA trees colored by AF3 RMSD relative to the 2021 reference for the full HA (A) and RBS-only region (B). (C) AF3 RBS geometry map showing joint variation in Δd130–190 and Δd150–220; points are colored by collection year and shaped by lineage (Lineage A, circles; Lineage B, squares). (D) Variant-level summary of apex and flank distance shifts (relative to 2021 reference) for key variants. (E–G) Trees colored by AF3-derived RBS distance shifts for Δd130–190 (E), Δd130–220 (F), and Δd150–220 (G), with blue-to-red indicating lower-to-higher values. For panels with diversity traces, the lower plots show time-resolved effective-bin diversity; grey shading indicates low-sample intervals, and the dashed vertical line marks the cattle-introduction period.



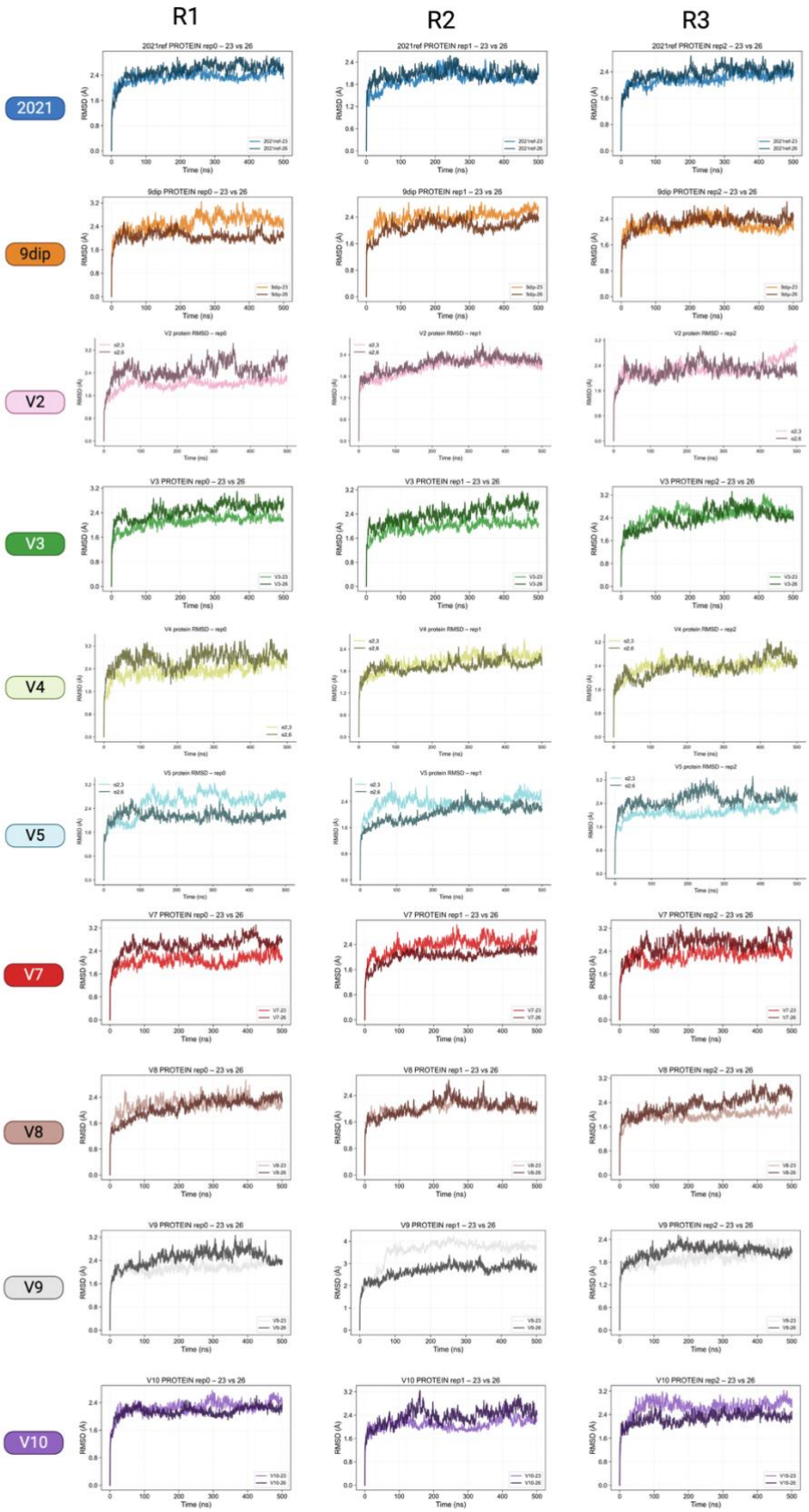
B

FlexddGwCR

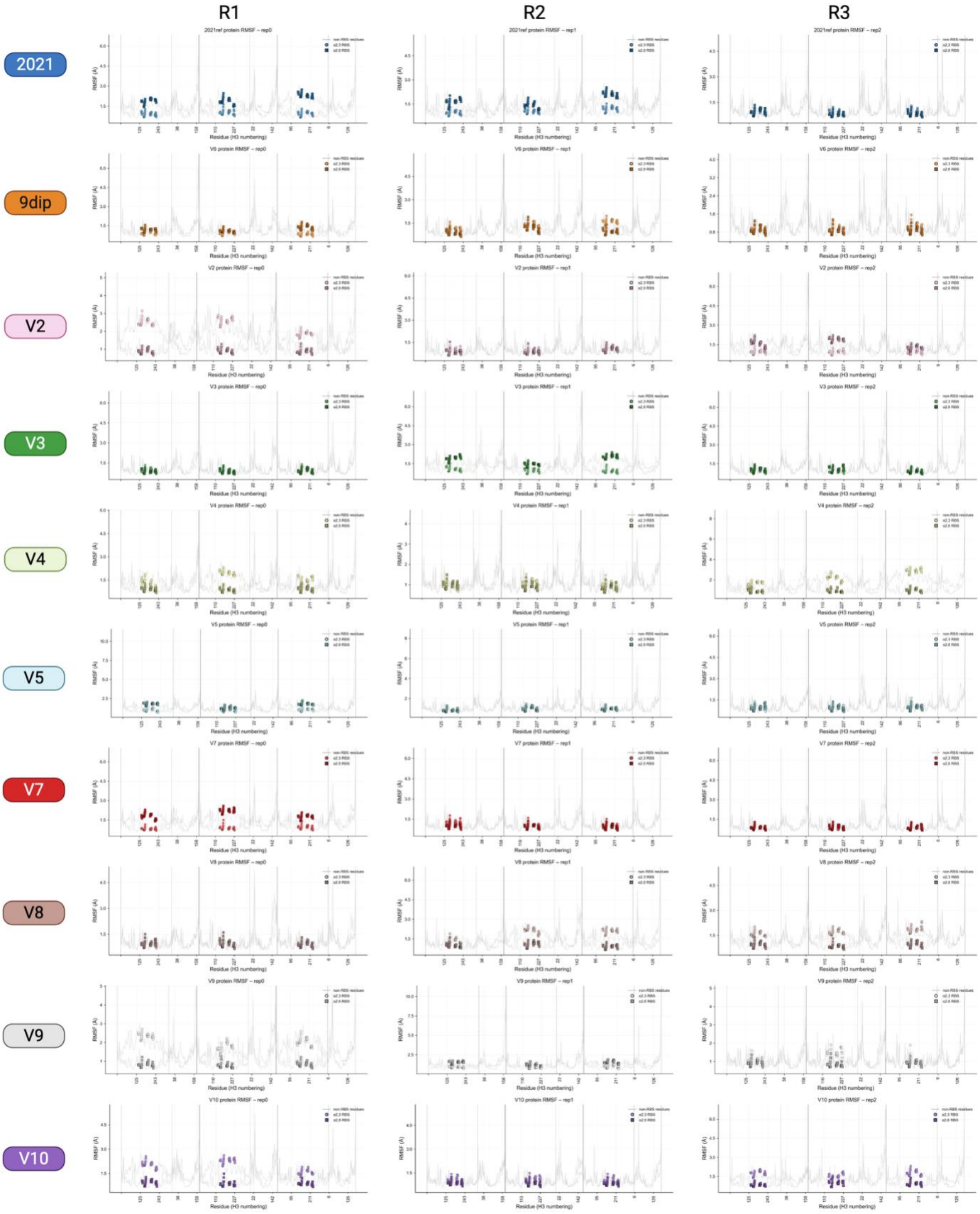
	Method	WT	Q226L	G228S	E190D	N224K + Q226L	N224K + Q226L + G228S
α23 sialoside	Rosetta	-	-1.460	3.063	4.512	2.614	3.902
	Wetlab (Lin *et. al.* 2024)	Strong	None	None	None	None	Weak
α26 sialoside	Rosetta	-	-8.622	-0.008	0.008	-3.693	-4.441
	Wetlab (Lin *et. al.* 2024)	None	Strong	None	None	Strong	Strong



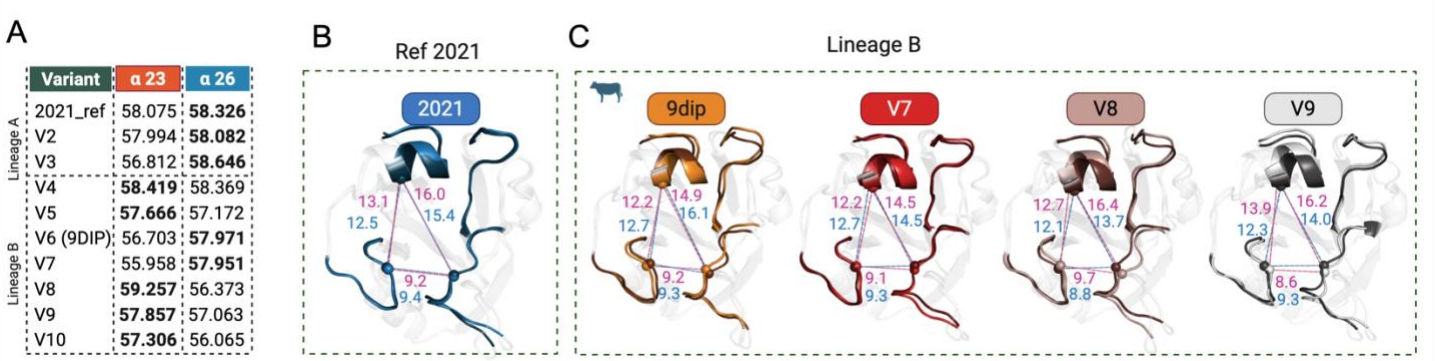
Supplementary Fig.8. Benchmarking Rosetta ΔΔG protocols and mapping HA stability and receptor binding on the phylogeny. **(A)** Scatter plot of alternative Rosetta ΔΔG protocols, showing RMSE (x-axis) and Pearson correlation (y-axis) for a curated HA benchmark set; each point corresponds to one protocol/feature combination. **(B)** Matrix summarizing predicted Rosetta ΔΔG values from the FlexddG with Cart pre-relaxation protocol for classic receptor-binding mutations (single and combined) with α2,3 and α2,6 sialosides, displayed alongside qualitative experimental annotations from Lin et al. 2024. **(C-D)** Time-scaled HA phylogeny of U.S. clade 2.3.4.4b sequences colored by predicted folding stability cartesian ΔΔG, and foldX. **(E–F)** Phylogeny colored by FlexddGwCR binding ΔΔG for α2,3 sialoside (E) and α2,6 sialoside (F), respectively, using a diverging color scale for negative and positive values (blue for negative and red for positive ΔΔG value). Lower panels show time-resolved effective-bin diversity for each metric; grey shading marks low-sample intervals, and the dashed vertical line marks the cattle-introduction period.



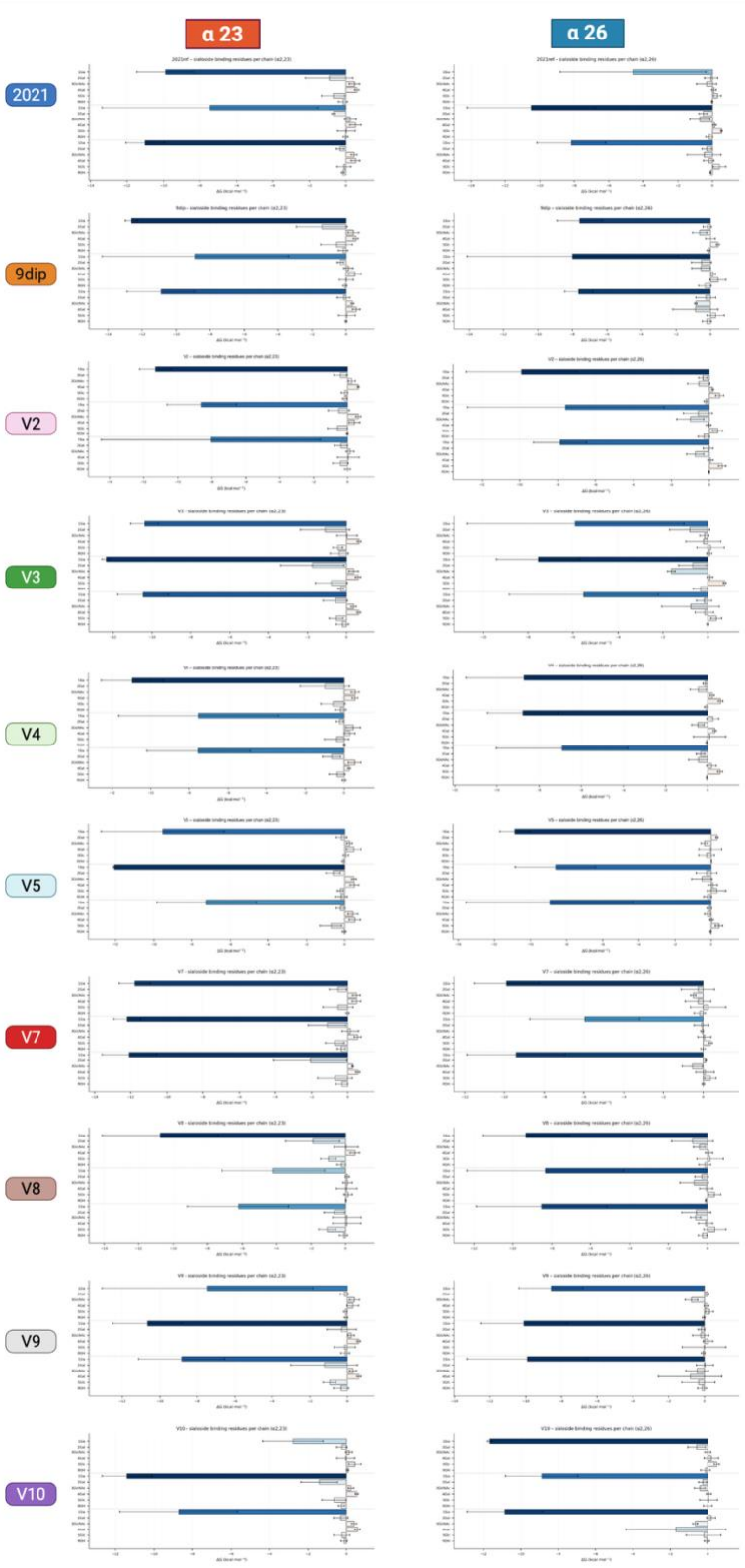
Supplementary Fig.9. Protein backbone stability across HA–sialoside MD simulations. For each variant backbone RMSD is plotted over time for three independent replicas, comparing simulations with α2,3 (light color) versus α2,6 sialosides (dark color) (paired traces in each panel).



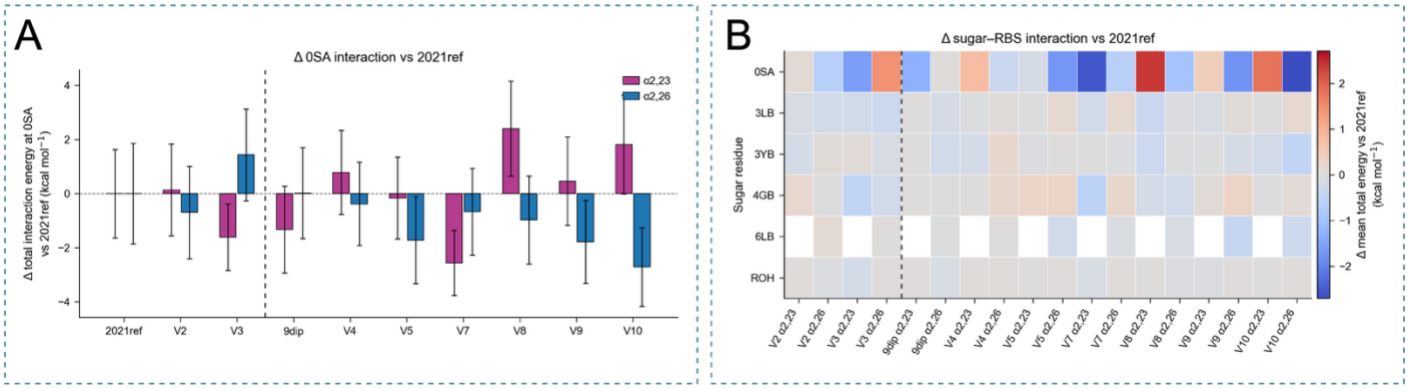
Supplementary Fig.10. Per-residue RMSF profiles across replicate HA–sialoside MD simulations. For each variant, residue-wise backbone RMSF is shown for three independent replicas, plotted along the HA sequence (H3 numbering), comparing simulations with α2,3 (circle) versus α2,6 linked sialosides (squares) (paired traces in each panel). Points represent per-residue fluctuations, with HA regions indicated by color along the x-axis.



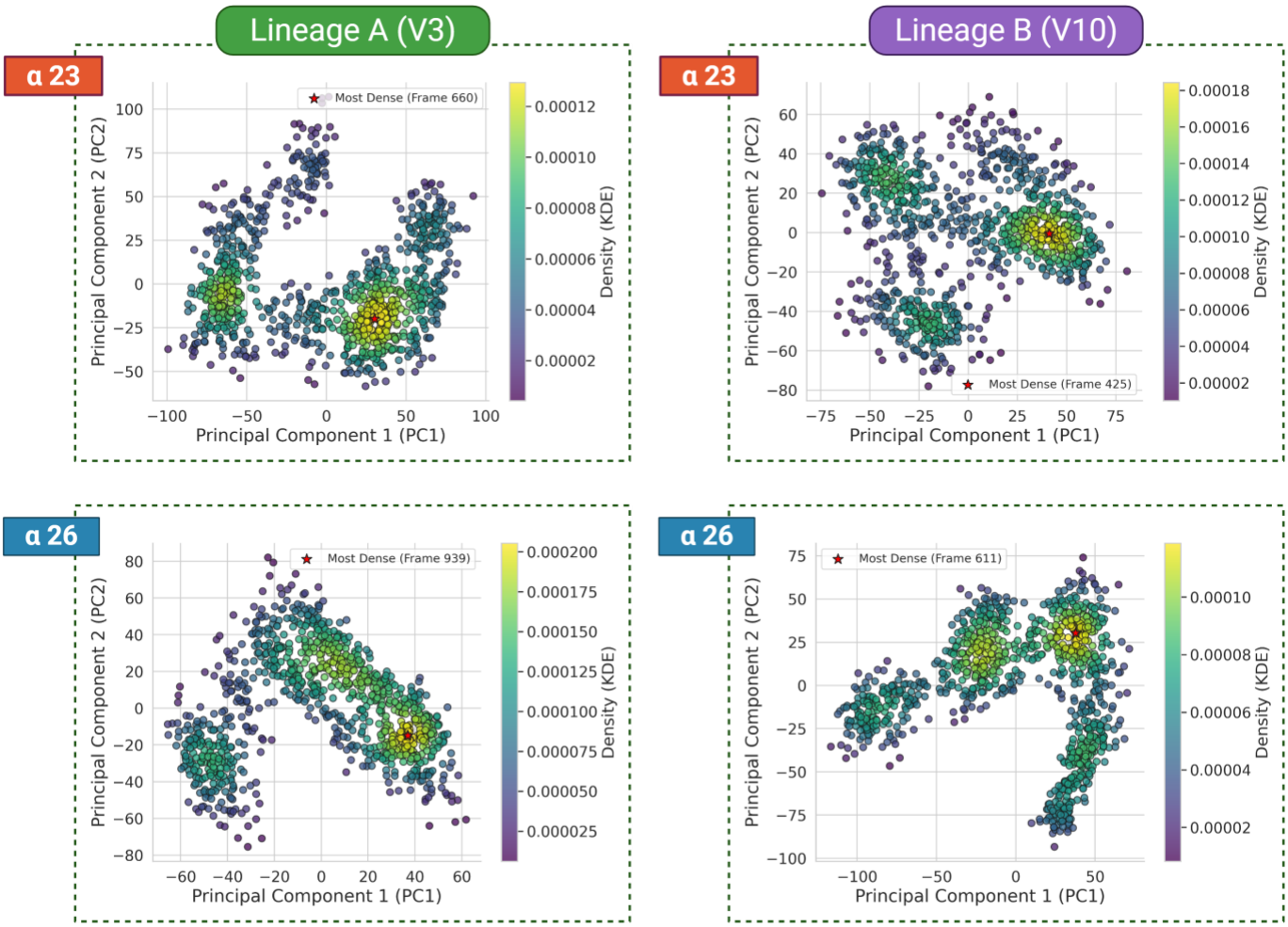
Supplementary Fig.11. Lineage-stratified RBS triangle geometry. (A) Mean RBS triangle area for all variants (B) geometry for the 2021 reference and (C) representative Lineage B variants (9DIP, V7, V8, V9), with edge distances annotated for α2,3 (pink) and α2,6 (blue).



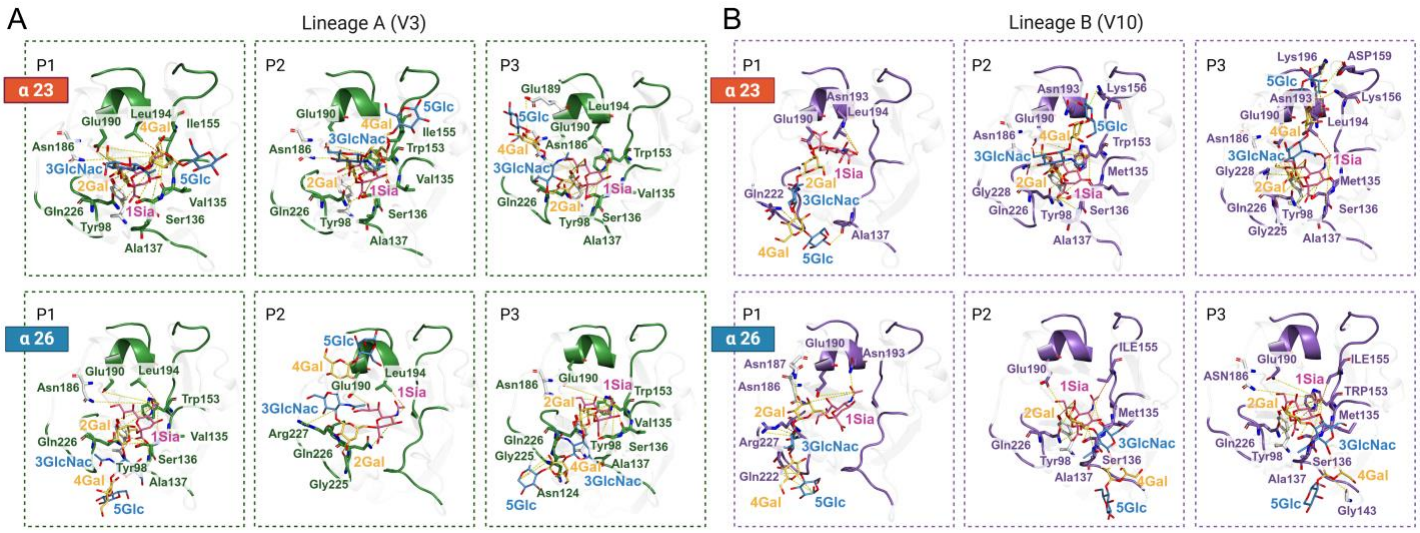
Supplementary Fig.12. Per-residue HA–sialoside binding energetics across clade 2.3.4.4b variants. ΔG contributions for individual sugar residues were computed from MD trajectories using MMGBSA and plotted. Horizontal bars show the mean ΔG (kcal mol^{-1}) for each sugar, with error bars denoting the standard deviation across replicates. Within each protomer, residues are ordered along the y axis from the reducing end (ROH) through 5Glc, 4Gal, 3GlcNAc, 2Gal to the terminal sialic acid (1Sia); dashed lines separate the three sialoside copies in the trimer. Colors report the magnitude and sign of the interaction energy (blue, more stabilizing; red, less stabilizing). This representation highlights consistent stabilization by the terminal 1Sia and neighboring galactose units, as well as variant-specific differences in how distal sugars contribute to overall receptor binding.



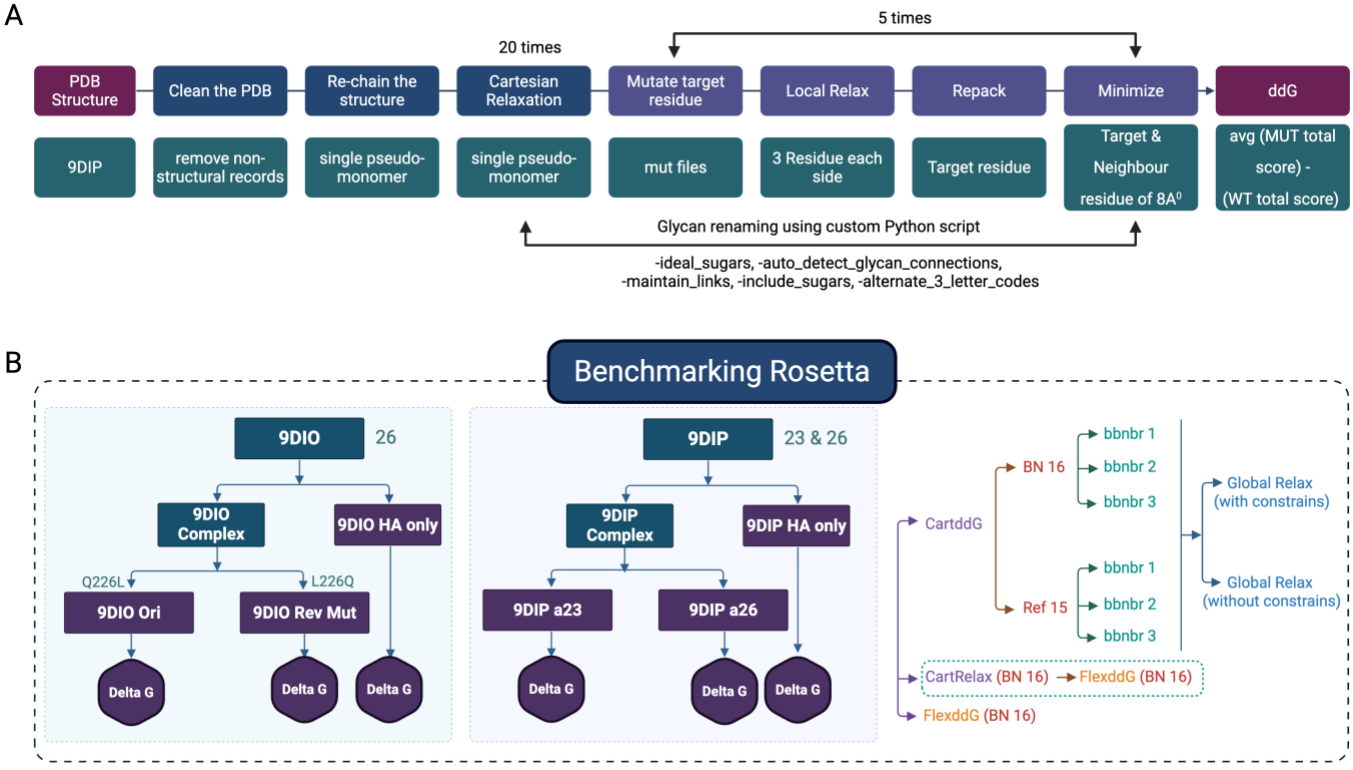
Supplementary Fig.13. Variant-specific redistribution of HA–sialoside binding energetics relative to the 2021 reference. Changes in glycan interaction energies are shown for the top clade 2.3.4.4b HA variants relative to the 2021 reference (2021ref) for α2,3- and α2,6-linked sialyllactose receptors; positive values indicate weaker, and negative values stronger, binding than 2021ref. **(A)** Δ total interaction energy of the terminal sialic acid (1Sia/0SA) for each variant, with α2,3 (magenta) and α2,6 (teal) linkages. **(B)** Heat map of per-sugar energy changes along the glycan (ROH→5Glc→4Gal→3GlcNAc→2Gal→1Sia), highlighting how different variants redistribute binding across individual sugar units.



Supplementary Fig.14. CA landscapes of MD trajectories for V3 and V10 HA-sialoside complexes. Principal component analysis of backbone motions for HA variants V3 (left) and V10 (right) bound to α2,3-linked (top) or α2,6-linked (bottom) sialosides. For each condition, individual MD frames are projected onto the first two principal components (PC1 and PC2), with points colored by kernel density (KDE) to indicate conformational sampling density and the corresponding density scale shown at right. The star in each panel marks the most densely populated conformational state (highest KDE frame).



Supplementary Fig.15. PLIP interaction fingerprints for Lineage A (V3) and Lineage B (V10) across protomers (P1–P3), shown for α2,3 (top) and α2,6 (bottom), with H bonds (yellow) and hydrophobic interactions (orange).



Supplementary Fig.16 (A). Rosetta cartddg2020 workflow for mutation scanning of the 9DIP HA trimer. (B) Benchmarking Rosetta with different approach to calibrate ΔΔG with the published wet-lab results from Lin et al [13].

Supplementary Table 1. Lineage-stratified phenotype summary.

Median and interquartile range summaries for phenotype metrics across Lineage A, Lineage B, cattle-associated Lineage B variants, and non-cattle Lineage B variants.

Metric	Lineage A (median, IQR)	Lineage B (median, IQR)	Lineage B, cattle-associated (median, IQR)	Lineage B, non-cattle (median, IQR)
pI	0 (-0.113 to 0)	0 (0 to 0.129)	0.1295 (0.129 to 0.478)	0 (0 to 0.129)
net_charge_pH7	-0.001 (-0.998 to 0.0015)	0 (0 to 0.999)	1 (0.9985 to 2.995)	0 (0 to 0.999)
gravy	0.006 (-0.001 to 0.012)	-0.006 (-0.012 to -0.001)	-0.008 (-0.01125 to -0.00375)	-0.005 (-0.012 to -0.001)
aliphatic_index	0.864 (0 to 1.552)	-0.176 (-0.542 to 0)	-0.365 (-0.512 to 0)	-0.176 (-0.5785 to 0)
pos_frac	0 (0 to 0.0002)	0 (0 to 0.0002)	0.0002 (0 to 0.0017)	0 (0 to 0.0002)
neg_frac	0 (0 to 0.0018)	0 (-0.0017 to 0)	-0.0017 (-0.00335 to -0.001125)	0 (-0.0017 to 0)
helix_frac	0.002 (0 to 0.004)	0 (-0.001 to 0)	-0.001 (-0.001 to 0)	0 (-0.001 to 0)
sheet_frac	0 (-0.002 to 0)	-0.002 (-0.002 to 0)	-0.002 (-0.00325 to -0.00075)	-0.002 (-0.002 to 0)
turn_frac	-0.001 (-0.001 to 0)	0 (0 to 0.002)	0.002 (0.002 to 0.00225)	0 (0 to 0.002)
pI_HA1	0.001 (0 to 0.001)	0 (0 to 0.589)	0.589 (0 to 1.437)	0 (0 to 0.589)
net_charge_pH7_HA1	0.001 (0 to 0.001)	0 (0 to 0.999)	0.999 (0 to 2.996)	0 (0 to 0.999)
gravy_HA1	0.002 (0 to 0.007)	0.006 (0 to 0.024)	0.024 (0.01925 to 0.025)	0.005 (0 to 0.024)
aliphatic_index_HA1	0.448 (0 to 2.197)	0.448 (0 to 1.748)	1.748 (1.748 to 1.748)	0.448 (0 to 1.748)
pos_frac_HA1	0 (0 to 0)	0 (0 to 0)	0 (0 to 0.0045)	0 (0 to 0)
neg_frac_HA1	0 (0 to 0)	0 (-0.0045 to 0)	-0.0045 (-0.009 to 0)	0 (-0.0045 to 0)
pI_HA2	0 (-0.069 to 0)	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)
net_charge_pH7_HA2	-0.001 (-0.999 to 0)	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)
gravy_HA2	0.001 (0 to 0.002)	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)
aliphatic_index_HA2	0.45 (0 to 0.45)	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)
pos_frac_HA2	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)
neg_frac_HA2	0 (0 to 0.0045)	0 (0 to 0)	0 (0 to 0)	0 (0 to 0)
af3_rmsd_all	0.233126 (0.185865 to 0.309713)	0.228197 (0.184606 to 0.354425)	0.248682 (0.199836 to 0.457294)	0.227917 (0.184381 to 0.352602)
af3_rmsd_rbs	0.036006 (0.0321222 to 0.045218)	0.0422783 (0.0370277 to 0.0603016)	0.047283 (0.0393051 to 0.0609374)	0.0421131 (0.0369302 to 0.0601461)
af3_delta_d130_190	0.00528188 (-0.00065067 to 0.0100474)	-0.00276524 (-0.009422 to 0.00383188)	-0.00906987 (-0.0142803 to -0.00541059)	-0.00261576 (-0.00939646 to 0.00396745)
af3_delta_d130_220	0.00577024 (-0.00148649 to 0.0167979)	-0.00655225 (-0.0165255 to 0.00154181)	-0.0188291 (-0.0253603 to -0.00896799)	-0.00628063 (-0.0159086 to 0.00194402)
af3_delta_d150_220	-0.0021881 (-0.00846605 to 0.00438041)	0.00553324 (-0.00472927 to 0.0279967)	0.0262426 (0.00211784 to 0.059623)	0.00528528 (-0.00490416 to 0.0269445)
cartddg_ddg_mean	1.464 (-1.6745 to 7.9045)	4.383 (-1.249 to 10.954)	9.1165 (4.94925 to 18.6125)	4.383 (-1.2605 to 10.7302)
cartddg_ddg_std	1.11 (0.64 to 1.851)	0.885 (0.3105 to 1.8235)	1.258 (0.61725 to 1.934)	0.881 (0.30175 to 1.8185)
ph_acid_mean_minus_neutral_298k	-0.167333 (-0.440154 to 0.0366144)	0.00283533 (-0.0111606 to 0.0341567)	0.0112956 (0.00104583 to 0.0120711)	0.00278 (-0.0121781 to 0.0391944)
flexddg_bind_23	-0.061 (-0.207 to 0.0855)	-0.01 (-0.15725 to 0.178)	0.047 (-0.15775 to 0.783)	-0.0115 (-0.15725 to 0.17375)
flexddg_bind_26	-0.039 (-0.194 to 0.1165)	0.0045 (-0.164 to 0.19925)	0.059 (-0.07775 to 0.9075)	0.0035 (-0.16525 to 0.1955)

Supplementary Table 2. Top ten HA variant identities, host composition and phenotype metrics.

The top ten HA amino-acid variant classes are summarized in manuscript variant order (V1-V10) across circulation, host composition, sequence-derived phenotypes, structural metrics, Rosetta estimates and MM/GBSA-derived receptor-engagement metrics.

Supplementary Table 2a. Variant identity and circulation summary.

Variant	No. sequences	Representative strain	Representative accession	First date	Last date	Span (days)	Dairy/ cattle (n)	Variant group	Key notes
V1	1767	A/AmericanWigeon/SouthCarolina/22-000345-001/2021	EPI_ISL_18133029	12/30/21	10/18/23	657	0	Avian backbone (early waves)	Wild-bird dominated backbone; spans 2021-12-30 to 2023-10-18; repeatedly seeds poultry outbreaks; rare wild-mammal spillovers.
V2	234	A/goose/Wyoming/USDA-010159-001/2022	EPI_ISL_18132890	4/4/22	12/28/22	268	0	Early transient (died out)	Early avian-dominated variant; spans 2022-04-04 to 2022-12-28; not detected after 2022, consistent with a transient lineage.
V3	1132	A/snowgoose/USA/012363-001/2025	EPI_ISL_19859615	7/2/22	3/2/25	974	3	Multi-host clade (non-cattle)	Largest non-cattle clade; spans 2022-07-02 to 2025-03-02; mostly avian but broad spillover (cats, wild mammals, humans); no sustained dairy-cow circulation.
V4	161	A/snowgoose/Minnesota/22-037628-005/2022	EPI_ISL_19605941	7/2/22	1/18/24	565	0	Pre-dairy avian variants	Avian-dominated pre-dairy variant; spans 2022-07-02 to 2024-01-18; no cattle; mainly wild birds.
V5	160	A/Swainson's hawk/Montana/23-029229-001/2023	EPI_ISL_18690651	9/20/23	7/2/24	286	0	Pre-dairy avian variants	Avian-dominated variant emerging pre-dairy; spans 2023-09-20 to 2024-07-02; no cattle; occasional wild-mammal detections.
V6	1216	A/dairycow/Texas/24_009775-001/2024	EPI_ISL_19094585	12/19/22	3/2/25	804	870	Dairy/cattle-associated sub-tree	Core dairy/cattle-enriched variant; Texas dairy-cow detections start 2024-03-10; spans 2022-12-19 to 2025-03-02; retains avian component indicating bidirectional exchange.
V7	390	A/California/173/2024	EPI_ISL_19531296	7/2/24	3/2/25	243	299	Dairy/cattle-associated sub-tree	Dairy-branch subclade; spans 2024-07-02 to 2025-03-02; DM-dominant with clustered human cases.
V8	248	A/California/150/2024	EPI_ISL_19497982	7/2/24	3/2/25	243	173	Dairy/cattle-associated sub-tree	Dairy-branch subclade; spans 2024-07-02 to 2025-03-02; DM-dominant with multiple human cases.
V9	183	A/dairycow/USA/24_032286-001/2024	EPI_ISL_19548798	7/2/24	3/2/25	243	145	Dairy/cattle-associated sub-tree	Smaller dairy/cattle-enriched variant; spans 2024-07-02 to 2025-03-02; primarily dairy cows with occasional bird detections.
V10	366	A/dairycow/USA/036204-001-R2/2024	EPI_ISL_19859521	7/2/24	3/2/25	243	344	Dairy/cattle-associated sub-tree	Near-exclusive dairy-cow circulation; spans 2024-07-02 to 2025-03-02; limited spillback into birds.

Supplementary Table 2b. Host composition by variant.

Variant	Wild bird (n)	Domestic bird (n)	Domestic mammal/cattle (n)	Wild mammal (n)	Human (n)	Unknown bird (n)	Wild bird (%)	Domestic bird (%)	Domestic mammal/cattle (%)	Wild mammal (%)	Human (%)	Avian (%)	Mammal (%)
V1	1390	315	0	26	0	36	78.660	17.830	0	1.470	0	96.490	1.470
V2	177	46	0	1	0	10	75.640	19.660	0	0.430	0	95.300	0.430
V3	680	310	16	16	10	100	60.070	27.390	1.410	1.410	0.880	87.460	3.710
V4	147	11	0	0	0	3	91.300	6.830	0	0	0	98.140	0
V5	108	49	0	2	0	1	67.500	30.630	0	1.250	0	98.120	1.250
V6	177	104	894	35	4	2	14.560	8.550	73.520	2.880	0.330	23.110	76.730
V7	9	51	317	3	8	2	2.310	13.080	81.280	0.770	2.050	15.380	84.100
V8	0	64	175	0	9	0	0	25.810	70.560	0	3.630	25.810	74.190
V9	12	18	152	0	0	1	6.560	9.840	83.060	0	0	16.390	83.060
V10	8	14	344	0	0	0	2.190	3.830	93.990	0	0	6.010	93.990

Supplementary Table 2c. Dominant sampled species by variant.

Variant	Top sampled species
V1	turkey(339); chicken(216); SnowGoose(104); canadagoose(61); Backyardbird(60)
V2	canadagoose(61); chicken(37); mallard(21); baldeagle(13); americanwhitepelican(10)
V3	chicken(306); turkey(153); canadagoose(120); red-tailedhawk(58); mallard(48)
V4	snowgoose(22); greathornedowl(17); redtailedhawk(14); bluewingedteal(13); gadwall(12)
V5	chicken(43); turkey(32); cacklinggoose(11); gadwall(10); Canadagoose(8)
V6	dairycow(851); chicken(99); turkey(52); cat(21); mallard(21)
V7	dairycow(299); chicken(51); cat(18); California(8); turkey(4)
V8	dairycow(173); chicken(64); California(9); cat(2)
V9	dairycow(145); chicken(14); turkey(11); cat(7); emu(4)
V10	dairycow(344); chicken(14); turkey(8)

Supplementary Table 2d. Sequence-derived biophysical metrics.

Variant	Net charge pH 7 (HA)	Net charge pH 7 (HA1)	Net charge pH 7 (HA2)	pI (HA)	pI (HA1)	pI (HA2)	GRAVY (HA)	GRAVY (HA1)	GRAVY (HA2)	Aliphatic index (HA)	Aliphatic index (HA1)	Aliphatic index (HA2)
V1	0	0	0	0	0	0	0	0	0	0	0	0
V2	-0.001	0	-0.001	0	0	0	0.001	0	0.002	0	0	0
V3	-0.998	0.001	-0.999	-0.113	0.001	-0.069	0.012	0.007	0.001	1.552	2.197	0.450
V4	0	0	0	0	0	0	-0.013	0	0	-0.688	0	0
V5	0	0	0	0	0	0	-0.007	0.024	-0.009	-0.688	1.748	-1.757
V6	0	0	0	0	0	0	-0.004	0.024	0	0	1.748	0
V7	0.999	0.999	0	0.129	0.589	0	-0.003	0.025	0	0	1.748	0
V8	0.999	0.999	0	0.129	0.589	0	-0.007	0.025	0	-0.512	1.748	0
V9	2.995	2.996	0	0.478	1.437	0	-0.004	0.024	0	0	1.748	0
V10	2.995	2.996	0	0.478	1.437	0	-0.008	0.024	0	-0.512	1.748	0

Supplementary Table 2e. Amino-acid composition and secondary-structure metrics.

Variant	Positive fraction (HA)	Positive fraction (HA1)	Positive fraction (HA2)	Negative fraction (HA)	Negative fraction (HA1)	Negative fraction (HA2)	Helix fraction	Sheet fraction	Turn fraction
V1	0	0	0	0	0	0	0	0	0
V2	0	0	0	0	0	0	0	0	0

Variant	Positive fraction (HA)	Positive fraction (HA1)	Positive fraction (HA2)	Negative fraction (HA)	Negative fraction (HA1)	Negative fraction (HA2)	Helix fraction	Sheet fraction	Turn fraction
V3	0	0	0	0.002	0	0.004	0.004	0	-0.001
V4	0	0	0	0	0	0	-0.001	-0.002	0
V5	0	0	0	0	0	0	-0.001	-0.002	0
V6	0	0	0	0	0	0	0	-0.002	0
V7	0	0	0	-0.002	-0.004	0	0	-0.002	0.002
V8	0	0	0	-0.002	-0.004	0	-0.001	0	0.002
V9	0.002	0.004	0	-0.004	-0.009	0	0	-0.004	0.002
V10	0.002	0.004	0	-0.004	-0.009	0	-0.001	-0.002	0.002

Supplementary Table 2f. Structure, stability, Rosetta and MM/GBSA metrics.

Variant	AF3 RMSD HA	AF3 RMSD RBS	Stability cartddG mean	pH stability acid-neutral	Flex ddG binding alpha2,3	Flex ddG binding alpha2,6	MM/GBSA total delta alpha2,3	MM/GBSA total delta alpha2,6
V1	0	0	0	0	0	0	0	0
V2	0.223	0.027	11.548	-1.037	0.047	0	1.123	-8.553
V3	0.181	0.032	0.383	-0.440	1.250	0.050	-18.358	5.483
V4	0.130	0.037	4.414	0.001	-0.119	0.033	6.733	-2.320
V5	0.213	0.033	4.630	0.003	-0.248	0.306	-0.518	-15.749
V6	0.170	0.039	-1.351	0.001	-0.163	-0.086	-9.352	-3.963
V7	0.201	0.046	5.823	0.011	0.047	-0.068	-25.970	-4.914
V8	0.803	0.057	20.489	0.012	1.089	1.934	12.230	-14.098
V9	0.294	0.039	5.152	0.011	-0.105	-0.095	2.212	-18.539
V10	0.509	0.065	18.375	0.012	1.166	1.014	12.735	-28.695

Supplementary Data files

The following Supplementary Data files are submitted separately and provide the underlying sequence, metadata, phenotype, structural, Rosetta, molecular-dynamics and MM/GBSA data used in the manuscript and Supplementary Information.

1. reference_2021_AAseq.fasta. Amino-acid sequence of the 2021 reference HA used to anchor variant comparisons, mutation enumeration and reference-normalized phenotype calculations.
2. reference_2021_DNAseq.fasta. Nucleotide coding sequence of the 2021 reference HA used for alignment, variant mapping and phylogenetic reference coordinates.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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 α2,3 or α2,6 sialoside linkage.
13. 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.
14. 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.
15. 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.

16. S4.3_mmgbsa_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.
17. 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.
18. 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.
19. 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.
20. rosetta_benchmarking_all_predictions.csv. Predicted and experimental ΔΔG values for benchmark receptor-binding mutations across Rosetta protocols and sialoside linkages.
21. 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.