Supplementary Material: Different antigenic distance metrics generate similar predictions of influenza vaccine response breadth despite low correlation

Table of contents

1. Reproducibility instructions

TODO write this once we have everything on a public repo.

2. Extended Methods

2.1 Antigenic distance calculation

We calculated four different antigenic distance metrics for our study. In this section, we walk through how each method is calculated. Note that we only considered pairwise distances between strains of the same subtype. So we only computed distances between two A(H1N1) strains, between two A(H3N2) strains, or between two influenza B strains, we did not compute distances between A(H1N1) or A(H3N2) strains or between any A and B strains. However, since the two B lineages are quite similar and our panel included pre-divergence influenza B strains, we performed pairwise comparisons of all influenza B strains.

Temporal distance is the absolute value of the difference in the years of isolation between the two strains. For example, the difference between A/H1N1/California/09 and A/H1N1/Michigan/15 would be $|2015 - 2009| = 6$.

Dominant p-Epitope distance is the maximum length-normalized Hamming distance across the five major epitope sites on the HA head. After aligning the HA amino acid sequences for all of the strains, we removed the signal peptides from the sequences and used the previously identified epitope site locations for influenza A (1) and influenza B (2). Working pairwise with the sequences, we concatenated the residues for each epitope and calculated the Hamming distance between each epitope, and we divided the Hamming distance for a given epitope by the number of residues in that epitope. Then the p-Epitope distance for that pair of strains was the maximum of those epitope-wise distances.

Grantham's distance is a weighted distance based on biochemical properties that considers how different two differing residues at the same position are. We used Grantham's substitution matrix (3) to assign a value to each residue site between two sequences, based on the transition between amino acids. More different transitions are given higher weights. Then, for each pair of sequences, we sum the weights for that pair and divide by the length of the sequence.

Finally, **cartographic distance** is the euclidean distance between strains on antigenic cartography map. We built our cartographic maps from the combined table of post-vaccination titer data in our study, treating all person-years as independent occurrences (there is no clear meaning for repeat measurements in a dimension reduction analysis). We used `Racmacs`, which implements metric multidimensional scaling, to create and optimize the cartographic map (4). All of our maps were two dimensional, and we selected the best fitting map from 25 distinct `Racmacs` runs with random initializations, where each initialization was allowed to perform up to 100 L-BFGS optimization runs to relax the initial MDS cartography. Multiple optimization runs is necessary because different initial conditions can lead to different maps (5). Combining multiple runs by applying a method like generalized Procrustes analysis is theoretically possible (simple averaging won't work

because rotation and scaling need to be taken into account) but has not yet been studied or published so we instead chose the one overall best run.

For our models, we only considered the antigenic distance between the assay strain and the vaccine strain of the same subtype for a given HAI assay. Some of the assay strains used were influenza B strains isolated before the Victoria/Yamagata lineage divergence. Because our main question was about the antigenic distance, we compared pre-divergence B strains to both the Yamagata and Victoria vaccine strains in our analyses. To facilitate fair comparisons across subtypes and antigenic distance metrics, we min-max normalized the antigenic distance measurements within each combination of influenza season, subtype, and metric. After normalization, the antigenic distance for homologous measurements was set to 0, and the antigenic distance for the most different assay strain used in a given season was set to 1, with all other antigenic distance values falling in this interval.

2.2 Causal modeling and model formulation

TODO: Need to update the models to be stratified by vaccine instead of by subtype!!!

While we do not claim that our estimates are causal, we employed a graphical causal model to formulate our statistical models. While all statistical models are a mix between practicality and the best possible model, we hope that by formalizing our thinking, our models will be robust and correctly answer our research questions.

Our original dataset contained one record per HAI assay, indicating the individual, season, study site, time point (pre- or post-vaccination), vaccine dose, and assay strain for each record. The data also included the following demographic variables: age, birth year, sex assigned at birth, and reported race/ethnicity. The study also provided a list of vaccine strains for each formulation (see the section on vaccine formulation for a complete list). Note that we only analyzed standard dose vaccine recipients in our analysis, so we do not discuss the vaccine dose further.

We built a causal model for the effect of antigenic distance as a directed acyclic graph (DAG). We include the following variables in our causal model: U, unobserved confounders that could be partially explained by nuisance variation, but are not directly explained in our model; r, the self-reported race/ethnicity; s, the sex assigned at birth; p, the pre-vaccination titer; a, the individual's age; and b, the individual's year of birth; sv, the vaccine strain (for a given subtype); and sa, the assay strain for a particular HAI assay. The causal model we selected is shown in Figure 1.

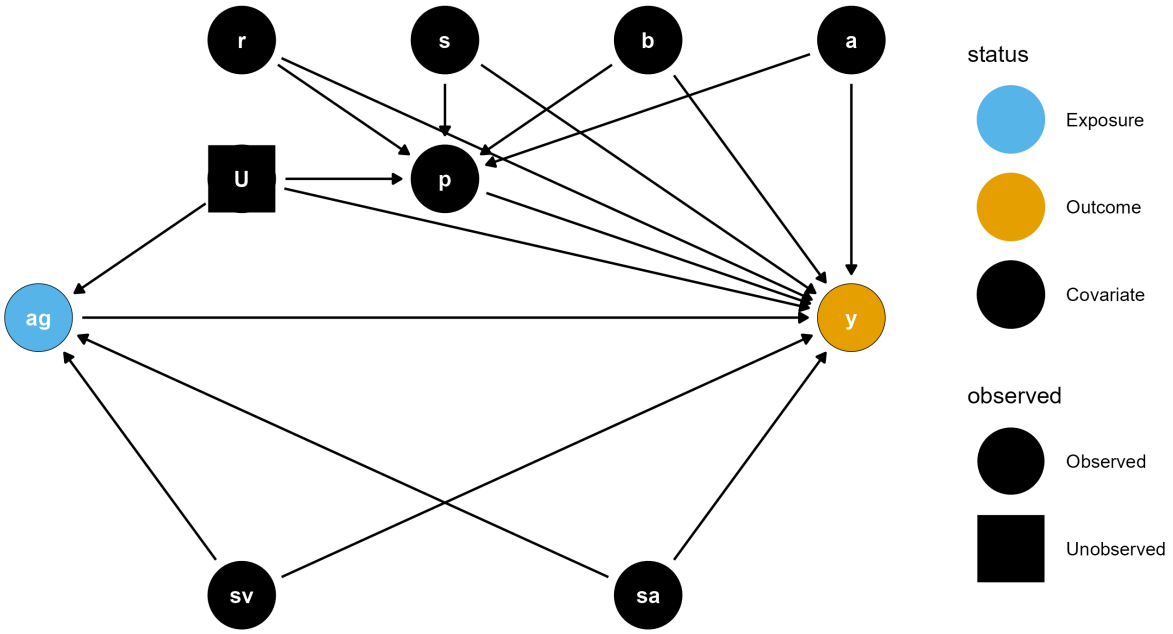

Figure 1: The graphical causal model for our research question represented as a DAG.

Under this causal model, the only confounders are the vaccine strain and assay strain, and any unobserved confounders. If we assume no unmeasured confounding, then the minimal sufficient adjustment set is only the vaccine strain and assay strain. However, our goal in this project was to analyze the effect of antigenic distance as a predictor without incorporating strain-specific effects. So, we stratified our models by vaccine strain (i.e., we fit all models separately for each vaccine component) and deliberately did not include a strain-specific effect.

To adjust for nuisance variation (potentially a source of unmeasured confounding), we included random effects to control for measurements at the same study site and on the same individual. Finally, we included specific ancestors of the outcome variable, which is not necessary to close backdoor paths, but does not improve bias and can improve the efficiency of the estimators of interest. We included pre-vaccination titer and age specifically in our model. In our previous work, we found that sex and race/ethnicity have minimal association with the observed HAI titers. Finally, our study was not specifically designed as an age-period-cohort analysis, and the age and birth year variables are highly correlated in our study (as one would expect). While our study has multiple longitudinal participants with different birth years, we elected to add only age to our model for simplicity. We included pre-vaccination titer in the model as-is, but since the age has a large range (from 11 to 65), we minmax scaled the age before using it in a model. Minmax scaling variables with large ranges can improve numerical stability of the model, but the model can still make predictions for any age.

Finally, we note that in some models it is also possible for cross-season differences to exist when the same vaccine strain was used for multiple years in a row. I.e., we might expect post-vaccination titers to change due to repeated usage of the same vaccine. However, since some of the vaccine strains were only used for one year before being replaced, this seasonal effect is not estimable in all of our models. Therefore, we decided not to include a seasonal effect in any of the models, especially since the effect of repeated usage of the same vaccine strain was not our primary research question.

2.3 Model implementation

We fit two models using `brms`, a generalized additive mixed model (GAMM) and a linear mixed model (LMM). The models were identical other than the specification for the effect of antigenic distance, so we will first describe the general parts of the model. Note that in the following mathematical descriptions, we adopt bracket notation rather than subscript notation following the convention of McElreath (6) due to the large number of subscripts in our model. That is, we use the notation $y[i]$ in place of the conventional y_i. We use subscripts to instead identify unique parameters. We also used the centered dot symbol $(\cdot)$ to avoid repetition when there are many valid arguments that would have the same right-hand side in a formula. For example, $\zeta[\cdot]$ indicates that all subscripts for ζ use the same equation.

We modeled our outcome (post-vaccination titer) as a Gaussian random variable, but due to the censored nature of our data we applied a censoring correction in the likelihood. Letting the outcome for a specific vaccine component be y, we assumed that

$$\begin{aligned} f(y_i \mid \mu[i], \sigma^2) &= \int_L^U \mathcal{N}\left(y[i], \mid \mu[i], \sigma^2\right) dy[i] \\ \sigma &\sim t^+(3,0,1) \\ i &= 1, \ldots, n \end{aligned}$$

where L and U are the lower and upper censoring bounds respectively (see the section on censoring bounds for details), $\mathcal{N}(\mu, \sigma^2)$ is the Gaussian (Normal) probability density function with mean μ and variance σ^2, $t^+(\nu, \mu, \sigma)$ is the location-scale half Student's t distribution with degrees of freedom ν, location parameter μ, and scale parameter σ. We chose a Student's t prior with $\nu = 3$ degrees of freedom because the distribution has fat tails, which allows the variance to be large if supported by the data, but we assume *a priori* that the distribution of the variance has a finite location and scale parameter (which is only the case when $\nu > 2$). Here, i is the index for the current data record and n is the total number of records in the dataset (i.e., each i indexes an HAI assay).

The model for the mean is shown below, including the priors for each parameter. For now, we represent the effect of antigenic distance as a function g, which we detail with its priors in the next section.

$$
\begin{aligned}
\mu[i] &= \beta_0 + u\left[1,\text{id}[i]\right] + u\left[2,\text{study}[i]\right] + g(\text{antigenic distance}[i]) + \\
&\quad \beta_p(\text{log pre-vaccination titer[i]}) + \beta_a(\text{scaled age}[i]) \\
\beta_0, \beta_p, \beta_a &\sim \mathcal{N}(0,1) \\
u[r,\cdot] &\sim \mathcal{N}(0, \omega[r]) \quad r = 1,2 \\
\omega[r] &\sim t^+(3,0,1)
\end{aligned}
$$

The priors follow the same formulation as before, but we chose Gaussian priors for the beta effects. Gaussian priors have flatter tails than Student's t priors, which provides a more regularizing effect for the beta parameters – that is, we presuppose that they are more likely to be close to zero, and our data needs to be strong enough to move the posterior distributions away from zero before we can make any conclusions.

The functional form of g is the only difference between the GAMM and the LMM. In the LMM, g takes a simple linear form:

$$
\begin{aligned}
g(\text{antigenic distance}[i]) &= \beta_d(\text{antigenic distance}[i]) \\
\beta_d &\sim \mathcal{N}(0,1)
\end{aligned}
$$

where the antigenic distance is minmax normalized for each model as described in the antigenic distance calculation section. For the GAMM, the function form of g is more complex. We modeled the antigenic distance effect using a thin-plate basis spline, which allows for the relationship to be curved in an arbitrary pattern, but constrains the fit so that rapid changes in the pattern are penalized and must be supported by data (7–11). The specific function form (using d to represent the normalized antigenic distance for readability) is

$$
\begin{aligned}
g(\text{d}[i]) &= \sum_{k=1}^{5} \gamma\,[k] \cdot \phi[k](\text{d}[i]) \\
\gamma[k] &\sim \mathcal{N}(0, \tau) \\
\tau &\sim t^+(3,0,0.25)
\end{aligned}
$$

where the $\gamma[k]$ are coefficients which are regularized to be similar via an adaptive prior and the $\phi[k]$ are thin-plate spline basis functions. Thin-plate splines use a low-rank approximation of the spline basis for computational efficiency, which can be tuned to balance between accuracy and efficiency. The maximal k (or size of the spline basis) we can choose is equal to the number of unique values of the predictor, so we chose $k = 5$, which was estimable across all of our antigenic distance metrics. We used Student's t priors for the adaptive prior on the variance of the spline coefficients so that the spline can be wiggly if supported by the data, but we chose a conservative hyperprior variance to constrain the spline towards being flat if the signal from the data is not strong.

2.4 Censoring bounds

HAI titer assays, like all serial dilution assays, produce censored data values. In fact, all values produced by an HAI assay are censored. We take this censoring into account in the likelihood of our model by integrating over the censoring bounds for a given data point y_i.

All serial dilution assays are censored – for the case of HAI, we assume that there is some latent, true dilution y_i^* which is the minimal dilution where hemagglutination is not observed. This is likely some decimal number, and we will never observe this true value. Instead, we chose a starting dilution, $y_{\min}$, which is 10 in our dataset. If we observe agglutination at this starting dilution, we say the value is below the limit of detection and it is recorded as 5 in our dataset. These values are left censored. In reality, we know that the latent agglutination dilution for an assay can be any value less than 10, i.e., our censoring bounds for these assays are $(0,10)$.

There is also a maxmimal dilution for the assay, $y_{\max}$, which was 20480 in our dataset. In practice, if researchers don't observe hemagglutination at any dilution, they can simply continue diluting the assay until they observe agglutination. However, a standard 96-well plate only has 12 columns, so most studies will report 20480 (the 12th serial dilution for an HAI assay starting at 10 and doubling each dilution). So these values are right censored, and the censoring bounds are $[20480, \infty)$. Note that the lower bound of the interval is included because the value *could* be exactly 20480 (though this occurs with probability zero for a continuous latent variable).

Finally, any other assay with a result between the limits of detection will also be interval censored, because we only observe certain dilutions. For example, if we observe inhibited hemagglutination at a dilution of 10, but agglutination occurs at a dilution of 20, we record the result as 10. However, we don't know that a dilution of 1:15 wouldn't care inhibition, so we only know that the latent dilution is in the interval $[10,20)$. Similarly for any value $y_{\min} < y < y_{\max}$, the latent dilution is in the interval $[y, 2y)$.

Converting to the log scale, the censoring bounds L and U that we refer to in the previous equations are as follows:

$$(L, U) = \begin{cases} (-\infty, y_{\min}) & y = y_{\min} \\ [y, y+1), & y_{\min} < y < y_{\max} \\ [y_{\max}, \infty) & y = y_{\max} \end{cases}.$$

3. Supplementary results

3.1 Annual Fluzone vaccine formulation

Table 1 shows the strains which were included in each season's formulation of the Fluzone vaccine. We only show the formulation for the standard dose (SD) vaccine, which differed from the HD vaccine formulation for several years.

Table 1: Strains used in the Fluzone standard dose vaccine formulation during each influenza season.

Season	A(H1N1)	A(H3N2)	B(Victoria)	B(Yamagata)
2013/14	CA/09	TX/12	—	MA/12
2014/15	CA/09	TX/12	—	MA/12
2015/16	CA/09	Switz/13	Bris/08	Phu/13
2016/17	CA/09	HK/14	Bris/08	Phu/13
2017/18	MI/15	HK/14	Bris/08	Phu/13

3.2 Annual heterologous strain panel

The strains used in each panel are shown in Table 2. A shaded cell with an X in it indicates that the strain indicated by the current row was used as part of the HAI panel in the season indicated by the current column.

Table 2: Heterologous strain panel used during each influenza season.

Subtype	Strain	2013/14	2014/15	2015/16	2016/17	2017/18
A(H1N1)	SC/18	X	X	X	X	X
	PR/34	X				
	Wei/43	X	X	X	X	X
	FM/47	X	X	X	X	X
	Den/57	X	X	X	X	X
	NJ/76	X	X	X	X	X
	USSR/77	X	X	X	X	X
	Bra/78	X			X	X
	CA/78		X	X		
	Chi/83	X	X	X	X	X
	Sing/86	X	X	X	X	X
	TX/91	X	X	X	X	X
	Bei/95	X	X	X	X	X

	NC/99	X	X	X	X	X
	SI/06	X	X	X	X	X
	Bris/07	X	X	X	X	X
	CA/09	X	X	X	X	X
	MI/15				X	X
A(H3N2)	HK/68	X	X	X	X	X
	PC/73	X	X	X	X	X
	TX/77	X	X	X	X	X
	MI/85	X	X	X	X	X
	Sich/87	X	X	X	X	X
	Shan/93	X	X	X	X	X
	Nan/95	X	X	X	X	X
	Syd/97	X	X	X	X	X
	Pan/99	X	X	X	X	X
	Fuj/02	X	X	X		
	NY/04	X	X	X	X	X
	Br/07	X				
	WI/05	X	X	X	X	X
	Uru/07		X	X	X	X
	Per/09	X	X	X	X	X
	Vic/11	X	X	X	X	X
	TX/12	X	X	X	X	X
	Switz/13	X	X	X	X	X
	HK/14		X	X	X	X
	Sing/16					X
B(Presplit)	Lee/40	X	X	X	X	
	MD/59		X	X	X	

	Sing/64		X	X	X	
B(Victoria)	Vic/87				X	X
	HK/01			X	X	X
	Mal/04			X	X	X
	Vic/06			X	X	X
	Bris/08			X	X	X
	CO/17			X	X	X
B(Yamagata)	Yam/88	X	X	X	X	X
	Harb/94	X	X	X	X	X
	Sich/99	X	X	X	X	X
	FL/06	X	X	X	X	X
	WI/10	X	X	X	X	X
	TX/11	X	X	X	X	X
	MA/12	X	X	X	X	X
	Phu/13	X	X	X	X	X

3.3 Strain names and abbreviations

Throughout the manuscript, we use abbreviated names for each strain. Table 3 shows the correspondening abbreviation for the full name of each strain.

Table 3: Full strain names and associated abbreviations for each strain used in the study.

Subtype	Strain name	Short name
A(H1N1)	A/H1N1/South Carolina/1/1918	SC/18
	A/H1N1/Puerto Rico/8/1934	PR/34
	A/H1N1/Weiss/1943	Wei/43
	A/H1N1/Fort Monmouth/1/1947	FM/47
	A/H1N1/Denver/1957	Den/57
	A/H1N1/New Jersey/8/1976	NJ/76
	A/H1N1/Ussr/90/1977	USSR/77
	A/H1N1/Brazil/11/1978	Bra/78
	A/H1N1/California/10/1978	CA/78
	A/H1N1/Chile/1/1983	Chi/83
	A/H1N1/Singapore/6/1986	Sing/86
	A/H1N1/Texas/36/1991	TX/91

	A/H1N1/Beijing/262/1995	Bei/95
	A/H1N1/New Caledonia/20/1999	NC/99
	A/H1N1/Solomon Islands/3/2006	SI/06
	A/H1N1/Brisbane/59/2007	Bris/07
	A/H1N1/California/07/2009	CA/09
	A/H1N1/Michigan 45/2015	MI/15
A(H3N2)	A/H3N2/Hong Kong/8/1968	HK/68
	A/H3N2/Port Chalmers/1/1973	PC/73
	A/H3N2/Texas/1/1977	TX/77
	A/H3N2/Mississippi/1/1985	MI/85
	A/H3N2/Sichuan/2/1987	Sich/87
	A/H3N2/Shandong/9/1993	Shan/93
	A/H3N2/Nanchang/933/1995	Nan/95
	A/H3N2/Sydney/5/1997	Syd/97
	A/H3N2/Panama/2007/1999	Pan/99
	A/H3N2/Fujian/411/2002	Fuj/02
	A/H3N2/New York/55/2004	NY/04
	A/H3N2/Brisbane/10/2007	Br/07
	A/H3N2/Wisconsin/67/2005	WI/05
	A/H3N2/Uruguay/716/2007	Uru/07
	A/H3N2/Perth/16/2009	Per/09
	A/H3N2/Victoria/361/2011	Vic/11
	A/H3N2/Texas/50/2012	TX/12
	A/H3N2/Switzerland/9715293/2013	Switz/13
	A/H3N2/Hong Kong/4801/2014	HK/14
	A/H3N2/Singapore/infimh-16-0019/2016	Sing/16
B(Presplit)	B/Lee/1940	Lee/40
	B/Maryland/1959	MD/59
	B/Singapore/1964	Sing/64
B(Victoria)	B/Victoria/02/1987	Vic/87
	B/Hong Kong/330/2001	HK/01
	B/Malaysia/27127/2004	Mal/04
	B/Victoria/326/2006	Vic/06
	B/Brisbane/60/2008	Bris/08
	B/Colorado/06/2017	CO/17
B(Yamagata)	B/Yamagata/16/1988	Yam/88
	B/Harbin/7/1994	Harb/94
	B/Sichuan/379/1999	Sich/99
	B/Florida/4/2006	FL/06
	B/Wisconsin/01/2010	WI/10
	B/Texas/06/2011	TX/11
	B/Massachusetts/02/2012	MA/12
	B/Phuket/3073/2013	Phu/13

3.4 Demographic information

A summary of the demographic information for the individuals included in our analysis is shown in Table 4, and includes information about their reported race/ethnicity, sex assigned at birth, age at first enrollment, and year of birth (see Supplement for detailed coding descriptions). The majority of participants identified their race as White or

Caucasian, and were assigned female at birth. All participants from the PA and FL study sites were adults, but the UGA study site also recruited teenagers, and all three study sites included elderly people over 65 years of age. Most participants returned to the study site in at least one subsequent year, contributing more than one person-year of data to the study.

Table 4: Demographic characteristics of the study participants. Summary statistics shown are count and column percent for sex, race, and contributed person-years; and median with range for age at first enrollment, birth year, and contributed HAI assays. Demographic variables were collected by a questionnaire from participants on the date they enrolled in a study season and received a vaccine. Coding details for the demographic variables are in the Supplement.

Characteristic	FL N = 241[1]	PA N = 133[1]	UGA N = 303[1]	Overall N = 677[1]
Sex Assigned at Birth				
Female	184 (76%)	93 (70%)	168 (55%)	445 (66%)
Male	57 (24%)	40 (30%)	135 (45%)	232 (34%)
Race/Ethnicity				
White	190 (79%)	70 (53%)	233 (77%)	493 (73%)
Black or American American	14 (6%)	52 (39%)	24 (8%)	90 (13%)
Other	12 (5%)	8 (6%)	33 (11%)	53 (8%)
Hispanic or Latino	24 (10%)	3 (2%)	13 (4%)	40 (6%)
Unknown	1 (0%)	0 (0%)	0 (0%)	1 (0%)
Age at First Enrollment	42 (20 - 80)	60 (26 - 81)	25 (12 - 83)	40 (12 - 83)
Year of Birth	1972 (1933 - 1996)	1954 (1932 - 1987)	1991 (1934 - 2006)	1975 (1932 - 2006)
Contributed HAI assays	85 (40 - 189)	94 (8 - 185)	48 (47 - 95)	52 (8 - 189)
Contributed person-years				
1	114 (47%)	44 (33%)	206 (68%)	364 (54%)
2	52 (22%)	31 (23%)	97 (32%)	180 (27%)
3	61 (25%)	32 (24%)	0 (0%)	93 (14%)
4	14 (6%)	26 (20%)	0 (0%)	40 (6%)

[1]n (%); Median (Min - Max)

Figure 2 shows a visualization of the collected pre-vaccination titers, and Figure 3 shows a visualization of the collected post-vaccination titers, ignoring all variables except for the assay strain.

Qualitatively summarizing the distribution of titers to all of the assay strains from plots alone is difficult, and the models in the main text are very helpful for understanding the variation in post-vaccination titers. However, we can make a few observations. Most people had some prior immunity (Figure 2) to the A(H3N2) strains which have circulated since the 80's or 90's, with protective (40 or greater) titers to the strains from the 2000's and onwards. However, most people only had protective titers to the two most recent

A(H1N1) strains, CA/09 and MI/15 which represent the 2009 pandemic lineage. Some people had immunity to older strains, but the difference was much more stark than for H3N2. Many people had prior immunity to all of the B strains we examined, and the median was 40 or greater for all of the B strains except MD/59.

Post-titers were, in general, higher (Figure 3). The two pandemic-like H1N1 strains showed a boost on average in the population, and there was noticeable back-boosting to some of the older H1N1 strains. Many of the H3N2 strains showed backboosting as well, although there was not much of a response to the oldest H3N2 strains which also had low pretiters. The median post-titers were above 40 for all of the B strains in our data, with Yamagata having the highest average titers, followed by Victoria and then the older lineages.

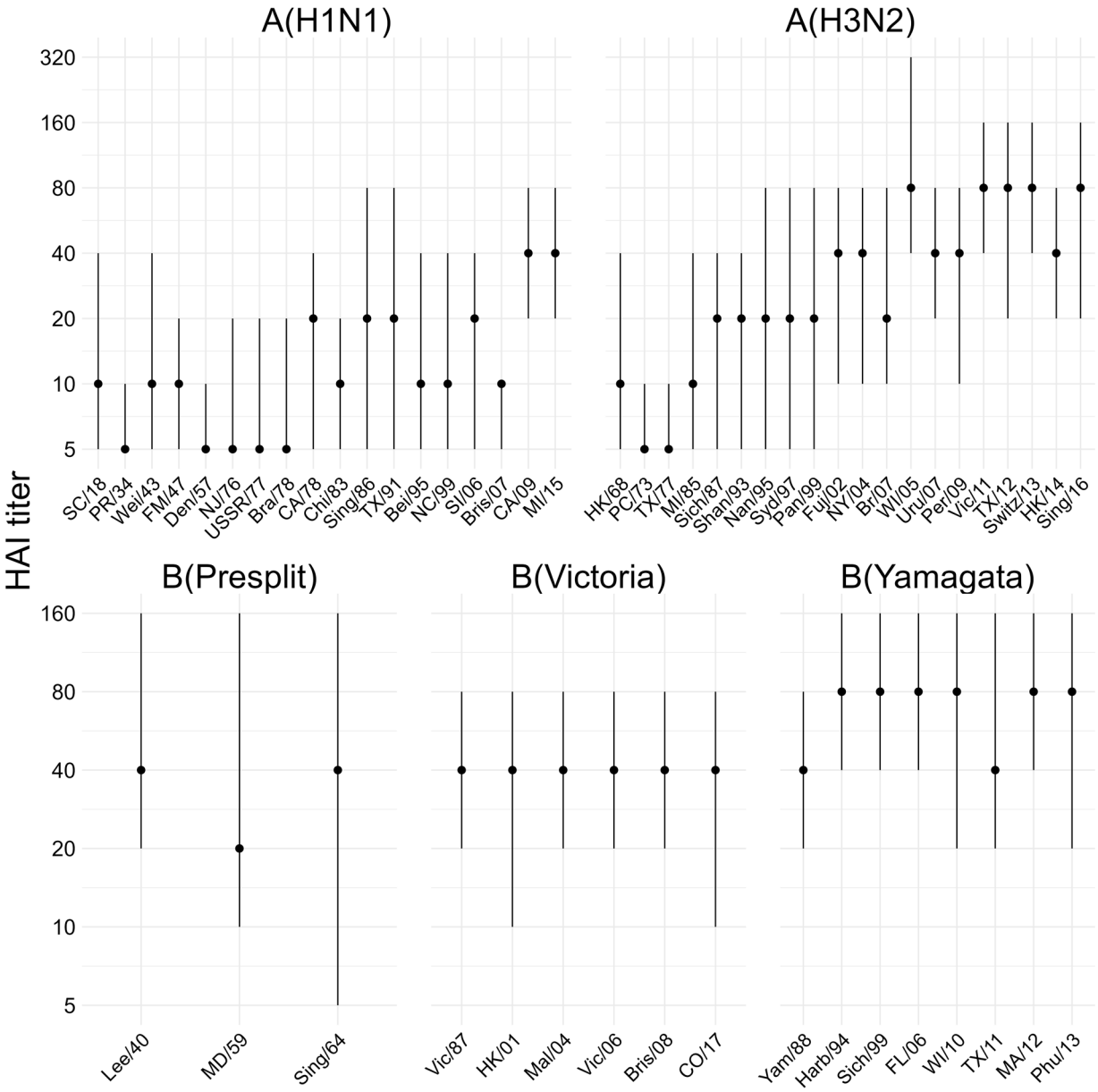

Figure 2: Pre-vaccination titers in our study to each of the assay strains. The point shows the median and the line shows the IQR.

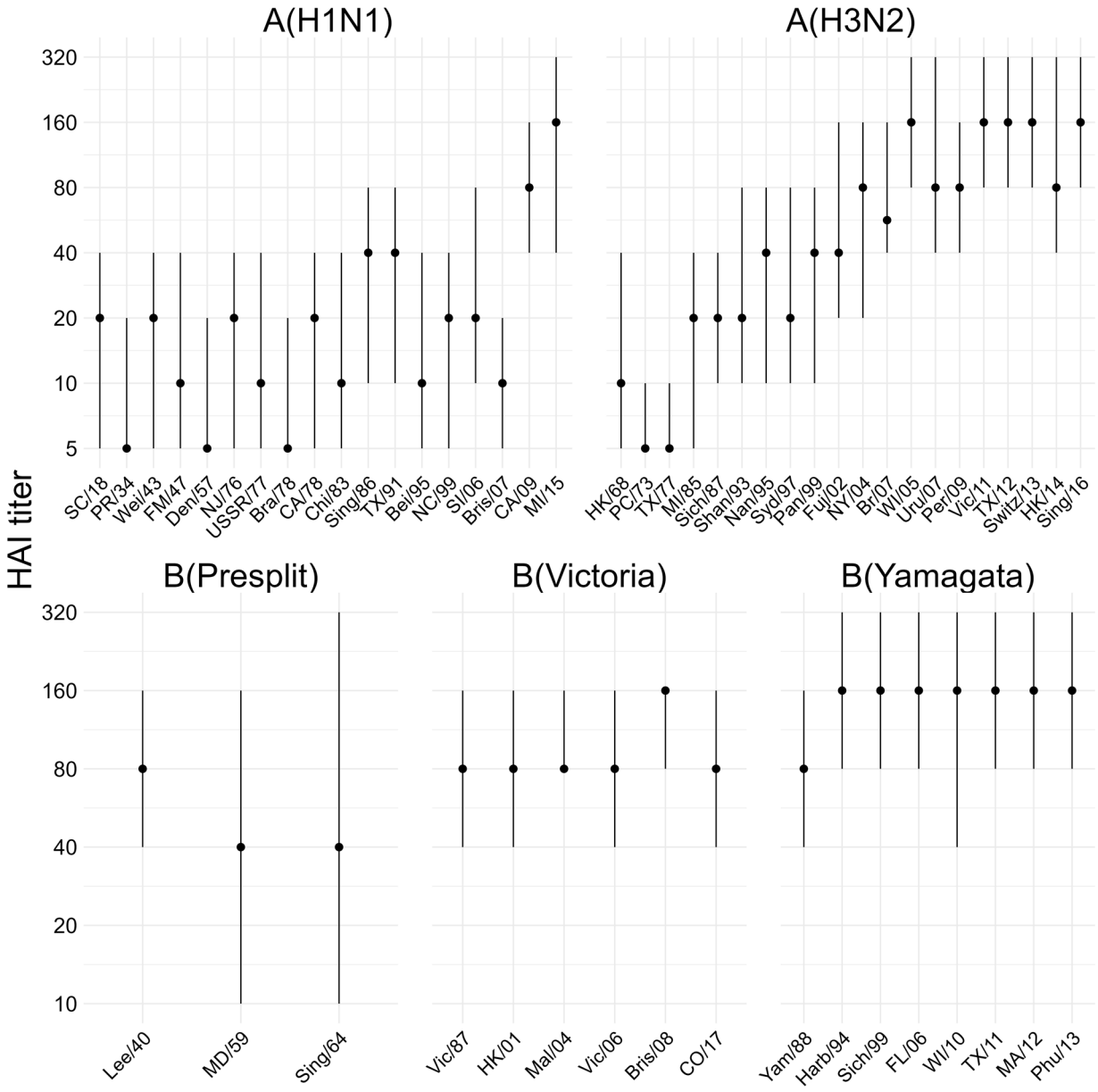

Figure 3: Post-vaccination titers in our study to each of the assay strains. The point shows the median and the line shows the IQR.

3.5 Metric agreement analysis

Before we built statistical models for the post-vaccination titer, we first performed a simple unadjusted analysis of the consistency (or agreement) between the antigenic distance measurements. As an omnibus test of agreement, we calculated the intraclass

correlation (ICC) across the four antigenic distance measurements, separately for each strain type. We used a Bayesian model with a fixed effect for antigenic distance metric and random intercepts for both assay strain and vaccine strain, and calculated the ICC as the ratio of variance explained by the assay and vaccine strain variance components to the total variation. The spearman correlations show in the main text can be viewed as a post-hoc analysisof the ICC which provide more information about specific comparisons.

Specifically, the model we fit for each subtype can be written as follows.

$$
\begin{aligned}
d[i] &\sim \mathcal{N}(\mu[i], \sigma^2) \\
\mu[i] &= \alpha_1 \cdot I(\text{method}[i] = \text{temporal}) + \alpha_2 \cdot I(\text{method}[i] = \text{p-Epitope}) + \\
&\quad \alpha_3 \cdot I(\text{method}[i] = \text{Grantham}) + \alpha_4 \cdot I(\text{method}[i] = \text{cartographic}) + \\
&\quad u\big[1,\text{assay strain}[i]\big] + u\big[2,\text{vaccine strain}[i]\big] \\
\alpha[k] &\sim t(3,0,5); \quad k = 1,2,3,4 \\
u[r,\cdot] &\sim \mathcal{N}(0, \zeta[r]); \quad r = 1,2 \\
\zeta[r] &\sim t^{+}(3,0,1) \\
\sigma &\sim t^{+}(3,0,1)
\end{aligned}
$$

We fit the model using Stan's NUTS sampler using 12 chains, each with 1000 warmup iterations and 1000 post-warmup sampling iterations and an adaptive delta of 0.99. Model diagnostics were all sufficient (data not shown, the model is easy to sample from and samples quickly). We then calculate the ICC as

$$\text{ICC} = \frac{\zeta_1^2 + \zeta_2^2}{\zeta_1^2 + \zeta_2^2 + \sigma^2},$$

over the posterior samples of all parameters. That is, the ICC represents the ratio of variance due to strain effects only to the total variance after controlling for fixed effects. In the psychometric literature, this is referred to as a one-way ICC for consistency – if the ICC is close to one, it means the variance from the random effects dominates the model. We summarized the ICC as the mean and 95% HDCI across the posterior samples.

The ICC was relatively low for all subtypes except A(H3N2), which had a moderate ICC (Table 5). The lower credibility limit included zero for all subtypes except A(H3N2), so despite the moderate point estimate for B(Yamagata) with a high upper limit, there was low consistency in antigenic distance measurements across methods. For A(H3N2), we observed moderate agreement across methods.

Table 5: Intraclass correlation (ICC) across all antigenic distance measurements, calculated separately for each subtype or lineage (strain type). The posterior distribution for each ICC was calculated as the ratio of variance components for vaccine strain and assay strain divided by the sum of all variance components, estimated with a Bayesian model. Numbers shown are the mean and 95% highest density credible interval (HDCI) of the posterior distribution of ICCs.

Strain Type	ICC
H1N1	0.08 (0.00, 0.23)
H3N2	0.34 (0.19, 0.52)
B-Yam	0.22 (0.00, 0.44)
B-Vic	0.03 (0.00, 0.14)

As a sensitivity analysis, we considered an alternative agreement statistic based on a different variance decomposition. We fit the same models as before, but then computed the variance of the posterior predictions for every point in the dataset without taking the random effects into account (the "fixed effects" predictions), i.e.

$$\sigma^2_{\text{FE}} = \text{Var}_{i=1}^{n}\left(\alpha\left[\text{method}[i]\right]\right),$$

where we choose the correct α parameter based on the method for dataset entry i (we omit writing all four alpha parameters and indicator functions for readability). Then, we compute the variance of the posterior predictions for each entry in the dataset taking the random effects and fixed effects into account:

$$\sigma^2_{\text{ME}} = \text{Var}_{i=1}^{n}\left(\alpha\left[\text{method}[i]\right] + u\left[1,\text{assay strain}[i]\right] + u\left[2,\text{vaccine strain}[i]\right]\right).$$

We can then compute an alternative agreement statistic as the variance ratio

$$1 - \sigma^2_{\text{FE}}/\sigma^2_{\text{ME}},$$

which will be close to one if the random effects dominate the prediction variance, or close to zero if the random effects have only a small contribution to the prediction variance. Table 6 shows our results using this metric. All of the results indicate low agreement but with a much higher uncertainty, and this metric is less charitable to the A(H3N2) consistency, although we observed strong pairwise correlations between all of the A(H3N2) metrics as shown in the main text.

Table 6: Prediction variance ratio across all antigenic distance measurements, calculated separately for each subtype or lineage (strain type). The posterior distribution for each ratio was calculated as one minus the ratio of the prediction variance ignoring random effects to the prediction variance including random effects, estimated with a Bayesian model. Numbers shown are the mean and 95% highest density credible interval (HDCI) of the posterior distribution of variance ratios.

Strain Type	PPD Ratio
H1N1	0.03 (-0.28, 0.30)
H3N2	0.21 (0.01, 0.39)

B-Yam	0.14 (-0.25, 0.48)
B-Vic	-0.05 (-0.75, 0.57)

3.6 Metric evenness and dispersion analysis

Since some of the antigenic distance metrics are more discrete than others, we calculated the gap standard deviation as a measure of evenness of distribution across each metric. The gap standard deviation is calcualted as the standard deviation of the consecutive differences in the sorted antigenic distance values for a given metric. That is, assume x, a vector of measurements from $i = 1, \dots, n$ is already sorted in increasing order so that $x_1 \leq x_2 \leq \cdots \leq x_n$. Then, the gap standard deviation is computed as

$$\begin{aligned} d_k &= x_{k+1} - x_k; \quad k = 1, \dots, i-1 \\ \bar{d} &= \frac{1}{n}\sum_{k=1}^{i-1} d_i \\ \sigma_{\text{gap}} &= \sqrt{\frac{1}{n-2}\sum_{k=1}^{i-1}\left(d_k - \bar{d}\right)^2}. \end{aligned}$$

For a random variable with a uniform distribution,

$$\lim_{n\to\infty} \sigma_{\text{gap}} = 0.$$

The different antigenic distance metrics also have different distributions in the set of observed variables. Rather than a uniform distribution of data points across distance space, each metric had gaps in the distribution of observed distances, which varied by metric and subtype (Figure 4 A). The two B lineages had much larger gaps due to the sparser historical panels. For influenza A, all metrics were more uniform for A(H3N2) than for A(H1N1), suggesting their different evolutionary patterns across the time spanned by the historical panel. Notably, while the temporal metric was the most uniform for all strains (an artifact of how the historical panel was chosen), the Grantham and p-Epitope metrics tend to discretize the number of potential distances and result in less uniformly distributed values for the historical panel used in our study.

We quantified the uniform spread of points for each antigenic distance metric and subtype using the gap standard deviation, where a higher gap standard deviation indicates more irregularity in the spacing of data points (see Supplement for details). Figure 4 B shows the estimated gap standard deviations. Both B lineages had higher gap standard deviations for all methods than either influenza A subtype. For A(H3N2), the gap standard deviations were similar across antigenic distance methods, and for A(H1N1) the differences were still small but larger than A(H3N2), representing the diversity of strains in the historical panel for type A influenza strains. The differences were much more noticeable for both B lienages, with Grantham distance having noteably higher gap standard deviation than the

other metrics for both influenza B lineages, indicating lower diversity in the normalized distance values.

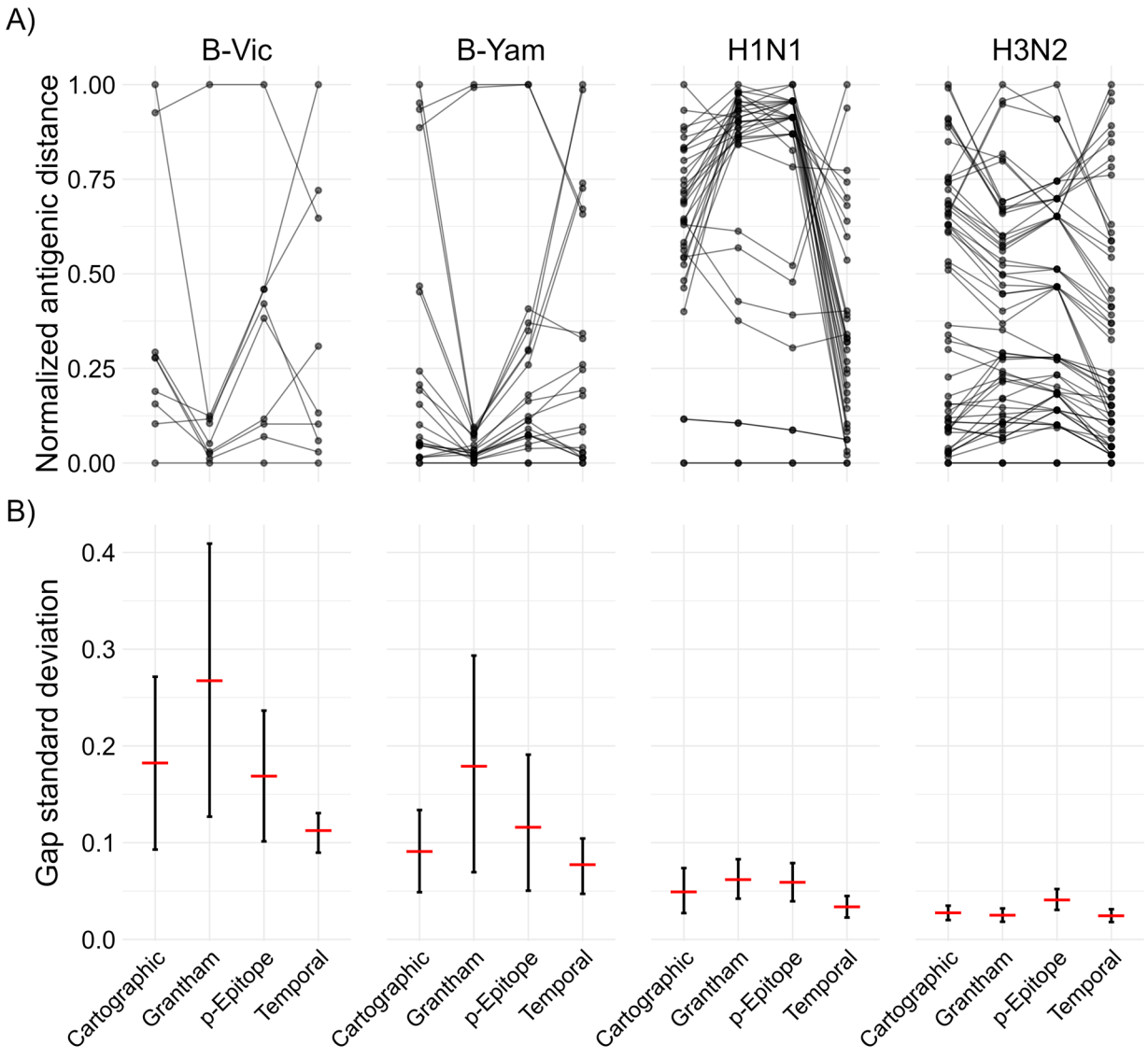

Figure 4: A) parallel coordinates plot showing how the estimated pairwise antigenic distances change for each of the antigenic distance metrics. Each line in the plot represents one vaccine strain and assay strain pair, and the connected points are the pairwise distance measured under each metric shown on the x-axis. When two lines cross, this indicates that two metrics assigned a different relative order to the pairwise combination. Note that Grantham and especially p-Epitope distances are integer-valued and concentrate measurements to specific points which potentially overlap (temporal distance is also integer valued but has enough spread to avoid a similar issue). B) The gap standard deviation (gap SD) for each subtype and antigenic distance metric. The posterior distribution of gap SDs was calculated using the bayesian bootstrap with

reweighting. The red horizontal bar shows the mean of the bootstrap posterior and the error bars show the 95% highest density credible interval (HDCI).

3.7 Model diagnostics

We examined the key model diagnostics for all of our models to ensure they converged. The main diagnostics with target criteria identified by the Stan development team (CITE THIS) are:

- $\hat{R}$, which measures chain mixing, should be < 1.01 for all parameters;
- Bulk and tail ESS, measures of the number of samples drawn if all of the samples were independent, should be greater than 100 times the number of chains;
- Number of divergent transitions should be less than 1% of samples;
- Number of treedepth exceedences should be less than 1% of samples;
- E-BFMI should be greater than 0.3 for all chains.

These diagnostics are presented in Table 7.

TODO: fit the new models and make sure to run them long enough to converge.

Table 7: Model diagnostics for the GAMMs and LMMs fit with each of the antigenic distance metrics. We show the total number of divergences out of the number of samples, and there were no treedepth exceedences for any of our models. For each model, we show the minimum ESS across all parameters, the minimum E-BFMI across chains, and the maximum R hat across all parameters.

Model	Metric	Num. Divergences	min ESS (tail)	min ESS (bulk)	min E-BFMI	max R_hat
GAMM	Cartographic	1 / 2400	93	55	0.468	1.222
LMM	Cartographic	1 / 2400	105	48	0.515	1.271
GAMM	Grantham	0 / 2400	20	51	0.504	1.236
LMM	Grantham	5 / 2400	104	41	0.518	1.334
GAMM	p-Epitope	22 / 2400	140	66	0.577	1.177
LMM	p-Epitope	2 / 2400	69	39	0.550	1.364

GAMM	Temporal	0 / 2400	28	46	0.402	1.286
LMM	Temporal	1 / 2400	81	37	0.507	1.400

We also examined trace plots of the parameters to ensure there were no obvious errors (and, in general, errors in the trace plots will be noticeable in the $\hat{R}$ statistic). We also examined the prior/posterior shrinkage and visually inspected prior/posterior plots. Since we have many models, each with hundreds or thousands of parameters, we did not include the plots here. We observed good values of shrinkage (far from 1, indicating a divergence away from the prior) for most parameters, with the exception of some highly constrained parameters, typically correlations and GAMM regularizing variance parameters. Some of the random effects for individuals had poor shrinkage as well, but overall the shrinkage for random effects and for the random effects variances was far from 1. Since the GAMM was not supported by ELPD anyways, we did not investigate prior sensitivity analysis further since all of the LMM parameters had good shrinkage. Therefore, we feel safe about our choice of regularizing priors and a prior sensitivity analysis would require extensive computational time without being useful.

3.8 Pointwise prediction comparisons

To examine the difference in predictions across each of the antigenic distance metrics, we computed the fold change in predicted post-vaccination HAI titer conditional on normalized antigenic distance and strain type for each unique pair of antigenic distance metrics. We visually inspected the conditional fold changes between metrics using a limit of agreement approach with a clinically defined threshold for whether the difference between predictions should matter, which is commonly defined as a 4-fold change for HAI measurements. We performed this fold change between predictions analysis for both the GAMM and LMM with each antigenic distance metric.

Figure 5 shows the prediction comparisons across antigenic distance metrics for each subtype using the LMMs. In contrast to our agreement analysis, where the H3N2 metrics showed the strongest agreement across metrics (and the highest pairwise correlations), H3N2 was the only strain with noticeable trends in the contrasts between metrics. In particular, all of the comparisons with p-Epitope for H3N2 had a noticeable trend – even though the mean fold change in predictions always stayed within the measurement error boundaries we set *a priori*, sometimes the credible interval did not fully cover the measurement error boundaries and there was a noticeable slope. These trends suggested that p-Epitope measurements underestimated the expected change in post-vaccination titer compared to Grantham and cartographic distance, while p-Epitope overestimated the difference compared to temporal methods. These results suggest that perhaps biochemical features like glycosylation sites or changes to the virus outside of the immunodominant epitope region are important, because these features are detected by cartographic and Grantham distance, but not by p-Epitope distance.

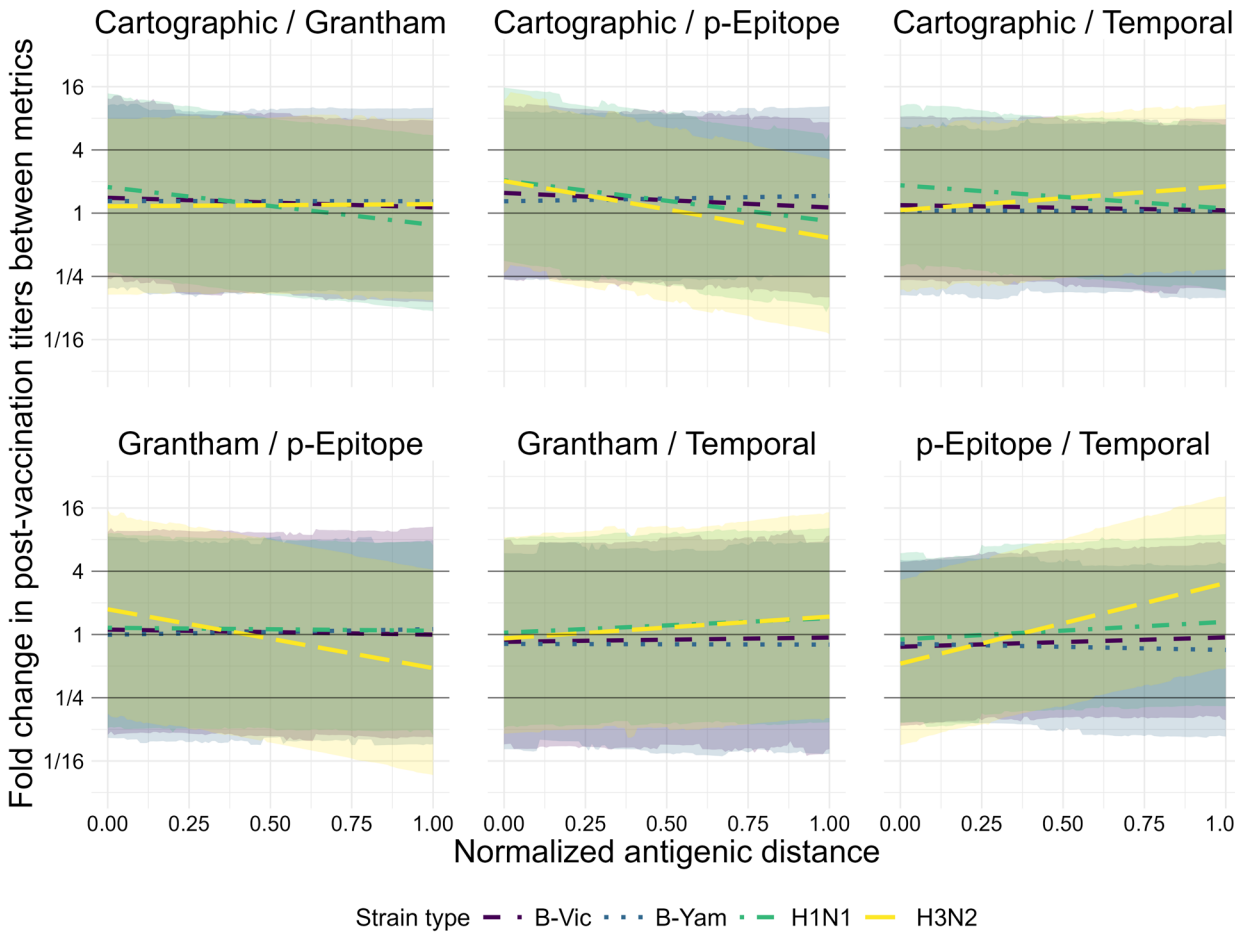

Figure 5: Pairwise comparisons of predictions (from the LMMs) between each unique set of two metrics. The y-axis shows the fold change in predictive titers between metrics, and the two metrics being compared in each subplot are shown as the subplot labels. Each line represents the predictions for the first metric in the pair at a given antigenic distance value divided by the predictions for the second metric in the pair. Color and linetype correspond to different strain types. The solid black lines on the plot are reference lines at a value of 1 for no effect, and at 4 and 1/4, effect values which would represent a clinically notable deviation in HAI predictions beyond what is expected from measurement error. Lines represent the mean of the posterior distribution of the contrast and the colored ribbons represent the 95% highest density credible interval (HDCI) for each strain type in each subplot.

Figure 6 shows the prediction comparisons across antigenic distance metrics for each subtype using the GAMMs. Even though the GAMM was not supported by our ELPD analysis, we used the GAMM for analyzing pairwise differences in predictions in case the nonlinear signal was biologically important with a weak signal. Unlike our simple correlation analysis, this analysis examines the predicted protection for an average individual exposed to an antigenically distant strain after vaccination, rather than only

taking antigenic distance into account. We saw that the fold change in predicted HAI titers was almost always less than four for every pairwise comparison between two metrics. A four-fold change in HAI titer is considered a clinically relevant difference between two measurements, so in almost every case we saw that changing the antigenic distance metric would not lead to a clinically relevant difference in predicted post-vaccination HAI titer. The primary exception was strain type A(H1N1), which exceed 40 at a few antigenic distance values for some of the pairwise comparisons (around a normalized antigenic distance of 0.25 for the cartographic/Grantham and Cartographic/p-Epitope comparisons, and around a normalized antigenic distance of 0.75 for the Grantham/temporal distance comparisons). Due to the large standard errors and the number of comparisons we make, we are comfortable attributing these fluctuations to measurement error, although the large variability across antigenic clusters for A(H1N1) strains (pdm-like vs. non-pdm-like) could contribute as well.

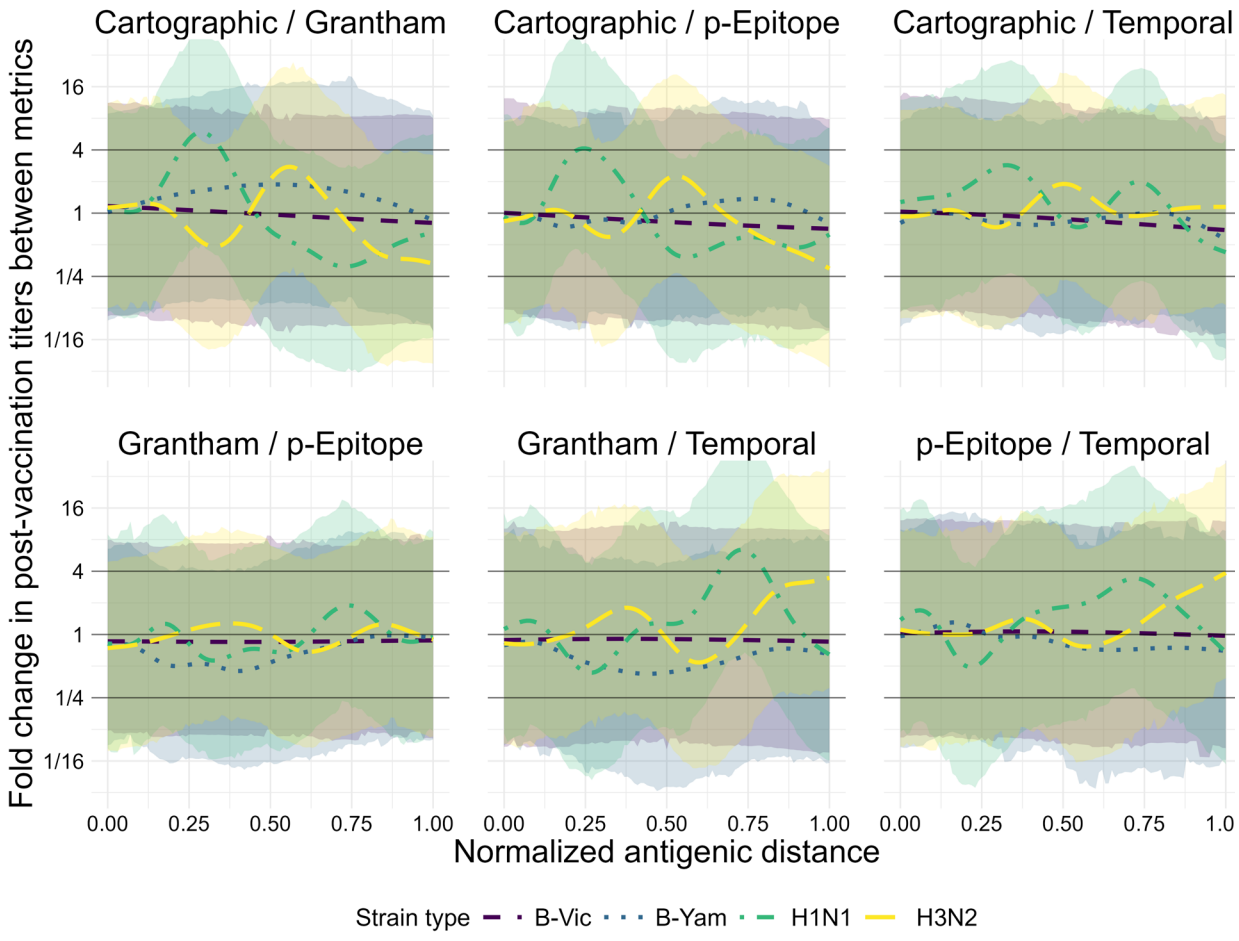

Figure 6: Pairwise comparisons of predictions (from the GAMMs) between each unique set of two metrics. The y-axis shows the fold change in predictive titers between metrics, and the two metrics being compared in each subplot are shown as the subplot labels. Each line represents the predictions for the first metric in the pair at a given antigenic distance value divided by the predictions for the second metric in the pair. Color and

linetype correspond to different strain types. The solid black lines on the plot are reference lines at a value of 1 for no effect, and at 4 and 1/4, effect values which would represent a clinically notable deviation in HAI predictions beyond what is expected from measurement error. Lines represent the mean of the posterior distribution of the contrast and the colored ribbons represent the 95% highest density credible interval (HDCI) for each strain type in each subplot.

However, the differences in comparisons for A(H3N2) was not completely trivial either. Figure 6 shows that for A(H3N2), the temporal distance overwhelming underestimates the fold change in predictions for the largest antigenic distances compared to both Grantham and p-Epitope measurements, with some interesting trends in the comparisons between cartographic distance as well. These results support our conclusion that further research into which of these metrics actually captures useful and interesting features is warranted, because it is difficult to tell whether we are capturing noise from our study or actual patterns that suggest different metrics are identifying different relevant characteristics of the viruses.

In both models, nearly all contrast predictions fall within the clinically irrelevant reference bounds, although the credible intervals for all predictions are wide because our bayesian models fairly account for many sources of uncertainty in the data. However, our results for the GAMM model suggest some interesting exceptions for the A(H1N1) strains that are likely related to the pandemic-like and non-pandemic-like cluster differences. Our results for the GAMM and LMM model for A(H3N2) seem to suggest that perhaps different metrics are picking up different relevant features, as we noted in the main text discussion.

4. References

1. Gupta V, Earl DJ, Deem MW. Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*. 2006;24(18):3881–3888.

2. Pan Y, Deem MW. Prediction of influenza B vaccine effectiveness from sequence data. *Vaccine*. 2016;34(38):4610–4617.

3. Grantham R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science*. 1974;185(4154):862–864.

4. Wilks S. Racmacs: Antigenic cartography macros. 2024.

5. Arhami O, Rohani P. TOPOLOW: A MAPPING ALGORITHM FOR ANTIGENIC CROSS-REACTIVITY AND BINDING AFFINITY ASSAY RESULTS. 2025;

6. McElreath R. Statistical rethinking: A Bayesian course with examples in R and Stan. Second edition. Boca Raton: CRC Press; 2020.

7. Wood SN. Generalized Additive Models: An Introduction with R, Second Edition. 2nd ed. New York: Chapman and Hall/CRC; 2017.

8. Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*. 2004;99(467):673–686.

9. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*. 2011;73(1):3–36.

10. Wood SN, N., Pya, et al. Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*. 2016;111:1548–1575.

11. Wood SN. Generalized additive models: An introduction with R. 2nd ed. Chapman and Hall/CRC; 2017.