PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A Bayesian approach to incorporate structural data into the mapping of genotype to antigenic phenotype of influenza A(H3N2) viruses

William T. Harvey[1¤*], **Vinny Davies**[2,3], **Rodney S. Daniels**[4], **Lynne Whittaker**[4], **Victoria Gregory**[4†], **Alan J. Hay**[4], **Dirk Husmeier**[3], **John W. McCauley**[4], **Richard Reeve**[1*]

1 Boyd Orr Centre for Population and Ecosystem Health, School of Biodiversity, One Health and Veterinary Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom, **2** School of Computing, College of Science and Engineering, University of Glasgow, Glasgow, United Kingdom, **3** School of Mathematics and Statistics, College of Science and Engineering, University of Glasgow, Glasgow, United Kingdom, **4** Worldwide Influenza Centre, The Francis Crick Institute, London, United Kingdom

† Deceased.
¤ Current address: Roslin Institute, University of Edinburgh, Edinburgh, United Kingdom
* wharvey@ed.ac.uk(WTH); richard.reeve@glasgow.ac.uk (RR)

OPEN ACCESS

Citation: Harvey WT, Davies V, Daniels RS, Whittaker L, Gregory V, Hay AJ, et al. (2023) A Bayesian approach to incorporate structural data into the mapping of genotype to antigenic phenotype of influenza A(H3N2) viruses. PLoS Comput Biol 19(3): e1010885. https://doi.org/10.1371/journal.pcbi.1010885

Editor: Rob J. De Boer, Utrecht University, NETHERLANDS

Received: March 28, 2022

Accepted: January 20, 2023

Published: March 27, 2023

Data Availability Statement: The HI data associated with this study are available at doi:10.5525/gla.researchdata.1405. Virus neutralisation data are available at https://www.crick.ac.uk/research/platforms-and-facilities/worldwide-influenza-centre/annual-and-interim-reports. HA sequence data are available at the Global Initiative on Sharing all Influenza Data (GISAID, https://www.gisaid.org). The described phylogenetic tree, sequence alignments, modelled datasets, scripts to generate phylogenetic and genetic variables, and to

Abstract

Surface antigens of pathogens are commonly targeted by vaccine-elicited antibodies but antigenic variability, notably in RNA viruses such as influenza, HIV and SARS-CoV-2, pose challenges for control by vaccination. For example, influenza A(H3N2) entered the human population in 1968 causing a pandemic and has since been monitored, along with other seasonal influenza viruses, for the emergence of antigenic drift variants through intensive global surveillance and laboratory characterisation. Statistical models of the relationship between genetic differences among viruses and their antigenic similarity provide useful information to inform vaccine development, though accurate identification of causative mutations is complicated by highly correlated genetic signals that arise due to the evolutionary process. Here, using a sparse hierarchical Bayesian analogue of an experimentally validated model for integrating genetic and antigenic data, we identify the genetic changes in influenza A (H3N2) virus that underpin antigenic drift. We show that incorporating protein structural data into variable selection helps resolve ambiguities arising due to correlated signals, with the proportion of variables representing haemagglutinin positions decisively included, or excluded, increased from 59.8% to 72.4%. The accuracy of variable selection judged by proximity to experimentally determined antigenic sites was improved simultaneously. Structure-guided variable selection thus improves confidence in the identification of genetic explanations of antigenic variation and we also show that prioritising the identification of causative mutations is not detrimental to the predictive capability of the analysis. Indeed, incorporating structural information into variable selection resulted in a model that could more accurately predict antigenic assay titres for phenotypically-uncharacterised virus from genetic sequence. Combined, these analyses have the potential to inform choices of

run described models are available at https://github.com/will-harvey/Flu_g2p_mapping. All other relevant data are within the manuscript and its Supplementary Information files.

Funding: This research was supported by the Medical Research Council (UK) under grant number MR/R024758/1 (WTH) and the Biotechnology and Biological Sciences Research Council (UK) under grants BB/L004828/1 (RR), BB/P004202/1 (RR) and BB/R012679/1 (RR) and by the programme grant to the Roslin Institute (award number BBS/E/D/20002173). The work performed at the London-based CC was supported by the Medical Research Council (1990-2014) and, subsequently, the Francis Crick Institute which receives its core funding from Cancer Research UK (FC001030), the Medical Research Council (FC001030) and the Wellcome Trust (FC001030). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

reference viruses, the targeting of laboratory assays, and predictions of the evolutionary success of different genotypes, and can therefore be used to inform vaccine selection processes.

Author summary

Mapping the impact of genetic changes on characteristics or traits is an important challenge in biology. The practical value of accurate genotype-to-phenotype mapping is clear in efforts to control human seasonal influenza viruses. These viruses, and particularly the A(H3N2) subtype, evolve rapidly with natural selection favouring variants possessing changes to the antigens recognized by the human immune system following prior infection or vaccination. This process of 'antigenic drift' necessitates global monitoring of the antigenic characteristics of the virus population and frequent vaccine updates. Viruses that are more closely related tend to be more antigenically similar. Consequently, in addition to the mutations causing antigenic drift, many antigenically neutral mutations are also highly correlated with antigenic drift. This complicates the task of accurately identifying the causative mutations and quantifying their antigenic impact. We present a modelling approach that attributes variation in antigenic assays to specific genetic changes while accounting for the evolutionary relatedness of viruses. We show that incorporating data on protein structure increases the accuracy of this process improving the reliability of genotype-to-phenotype mapping in this context. Precise genotype-to-phenotype mapping has the potential to improve understanding of the drivers of evolutionary success of emerging virus variants.

Introduction

Antigenic variation is a mechanism by which an infectious agent such as a virus or bacterium alters the proteins or carbohydrates exposed to the host immune system to allow escape from immunity conferred by prior infection with or vaccination against a related agent. Antigenic drift seen in influenza viruses is a prime example of this process. Human seasonal influenza epidemics are estimated to infect around 15% of the global population annually resulting in three to five million cases of severe illness and in the order of 290,000 to 650,000 deaths annually [1,2]. The lack of proofreading by the RNA polymerase contributes to a relatively high frequency of mutation across the genome. Influenza viruses evolve rapidly, with the mutation rate, or the accumulation of mutations during evolution, being particularly high in the genes encoding the surface glycoproteins where selection may favour genotypes encoding antigenic variants. Some antigenic variants may rise to dominance among circulating viruses due to strong immune-mediated positive selection favouring viruses that infect individuals previously immune due to prior infection or vaccination. Consequently, human seasonal influenza vaccines must be frequently updated to ensure the antigenic responses they elicit will be active against viruses in circulation.

Seasonal influenza epidemics are caused by viruses belonging to two influenza A subtypes, A(H1N1) and A(H3N2), and by influenza B viruses which are classified into antigenically distinct lineages, B/Victoria and B/Yamagata. Trivalent vaccines contain antigens based on the influenza A(H1N1) and A(H3N2) and the predominant influenza B lineage while quadrivalent compositions include antigens representative of both influenza B lineages. To monitor the

genetic and antigenic evolution of human influenza viruses, the WHO coordinates the Global Influenza Surveillance and Response System (GISRS) collaborating with academic scientists and national public health organisations [3,4]. Vaccine strain selection is based on the antigenic and genetic evolution of circulating influenza viruses throughout the year and recommendations are developed by representatives of the GISRS at twice-yearly vaccine composition meetings.

Within each of the influenza A subtypes and influenza B lineages that cause seasonal influenza epidemics, the global virus populations typically consist of several antigenically distinct groups of viruses. It takes around six months to develop, produce and deliver an updated influenza vaccine so strain selection decisions must be made up to nine months in advance of the period when influenza viruses will circulate in a forthcoming season. It is therefore necessary to understand the antigenic similarity of circulating viruses to current vaccine strains and vaccine candidates to predict which antigenic variants are most likely to circulate at high frequency in advance of a future influenza season, a task that benefits from predictive modelling [5]. Various approaches have demonstrated prediction of successful influenza lineages, from those emerging, with potential to inform vaccine virus selections [6–8]. One approach is to use the shape and branching pattern of haemagglutinin (HA) phylogenetic trees to track and extrapolate changes in genotype frequency [6]. Another approach is to predict lineage fitness using counts of amino acid substitutions inside and outside described antigenic sites as proxies for antigenic drift and reduced stability respectively [7]; an approach that can be adapted to incorporate data from assays used to measure antigenic drift. Both the accuracy of and the reliance on such predictive models depend on an understanding of the indirect link between genotype and reproductive fitness in a partially immune population, a relationship that is informed by the more direct phenotypic consequences of genetic changes.

The antigenic characterisation of circulating viruses is dependent upon haemagglutination inhibition (HI) and virus neutralisation (VN) assays, both of which are used to assess the antigenic similarity of a circulating test virus to a panel of reference viruses that includes previous and current vaccine viruses and other candidate vaccine viruses. The panel of reference viruses, and post-infection ferret antisera raised against them, are selected to represent the diversity of antigenic phenotypes observed over the most recent seasons. Various modelling approaches have used data from antigenic assays to quantify similarity [9], and to explore the relationship between genetic and antigenic evolution allowing predictions of antigenic relationships from sequence data [8,10–12]. A general challenge for modelling genotype-phenotype relationships is differentiating causative mutations from those that are non-causative and correlate with phenotypic changes due to genetic hitchhiking. Various phylogenetic comparative methods exist to account for shared evolutionary history of taxa when modelling quantitative traits, though these tend to focus on traits intrinsically associated with particular taxa rather than measures that relate to relationships between taxa, as is the case here when working with pairwise measures of antigenic similarity. However, by including terms that represent branches of the phylogenetic tree, it is possible to account for shared evolutionary history and to prevent false statistical support for genetic terms due to repeated measurements in the assessment of variation in VN titres for foot-and-mouth disease virus (FMDV) [13]. This approach was subsequently applied to influenza A(H1N1), demonstrated to preferentially identify genetic changes that correlate with antigenic changes in multiple locations across the phylogeny, and experimentally validated [12].

Comparing sequences of test and reference viruses, while accounting for phylogenetic correlations, we have previously identified substitutions responsible for antigenic evolution among human influenza A(H1N1) and avian influenza A(H9N2) viruses [12,14]. We then made this phylogenetically-aware model more statistically rigorous within a Bayesian

framework with antigenic determinants identified using 'spike-and-slab' priors [15], a method of variable selection (SABRE) demonstrated to outperform alternative approaches such as LASSO [16] and elastic net regularised regression [17] as well as our own previous work [12]. We further extended these approaches (eSABRE) by the inclusion of variables representing the underlying HI titre for each reference/test virus pair [18]. However, in order to allow us to fully characterise the algorithms involved, the data analysed in that study comprised only 43 viruses. These previous approaches can therefore be characterised as either applying methodologically-limited methods of model selection to datasets of scientific relevance [12], or the application of state-of-the-art Bayesian approaches to datasets of limited scale and relevance to the problem of antigenically characterising viruses [18].

A sequence-based model of virus fitness, grounded in an understanding of how amino acid variations affect fitness via changes to various aspects of virus phenotype, able to predict both the evolutionary success of existing genotypes and which unforeseen genotypes may be selected to emerge, would be a monumental achievement in the study of virus evolution. Challenging steps towards this goal include developing i) more accurate quantitative mappings of genotype to various phenotypes and ii) a quantitative understanding of how variations in various phenotypes contribute to evolutionary success in dynamic fitness landscapes. Addressing the former, our aim is to develop a practical tool for accurate, quantitative mapping of genotype to antigenic phenotype. Consequently, we seek to maximise the accuracy with which causative genetic differences are identified and quantified, rather than explicitly aiming to maximise predictive power. With this in mind, we present a hierarchical Bayesian model that uses using Bayesian stochastic search variable selection (BSSVS) to select genetic variables and apply it to a large set of antigenic data spanning 25 years and consisting of over 38,000 titres derived panels of antisera and contemporary circulating test viruses. The BSSVS approach, which aims to identify the combination of genetic variables best explaining observed antigenic variation, is performed within the model as it undergoes Markov Chain Monte Carlo (MCMC) sampling. Parameter estimates are made averaging over the best set of models, allowing uncertainty to be accounted for directly. A further advantage of a Bayesian approach is that existing knowledge can be used to define the prior distribution of parameters. These priors, together with the likelihood of the observed data given the statistical model, combine using standard Bayesian methods to form the posterior distribution of parameters. We show that information derived from measurements of solved protein structures can be used to shape prior distributions and improve the accuracy with which we can attribute changes in antigenic phenotype to causative amino acid substitutions. Finally, using Bayesian model averaging, where predictions are averaged over a range of the best supported models [19], these approaches show the ability to accurately predict antigenic relationships from genetic sequences.

Modelling approach

The approach presented infers the genetic determinants of antigenic evolution by attributing variation in antigenic assays to differences in the amino acid sequences of reference and test viruses, while accounting for both phylogenetic structure in the data and other non-antigenic factors that cause variation in titres. Using fixed quantities of reference and test viruses, commonly eight or four haemagglutinating units in a particular HI assay, titres are recorded as the reciprocal of the maximum dilution of antiserum raised to a particular reference strain that is able to inhibit agglutination of red blood cells by a test virus. Lower titres, expressed as fold-drops, therefore reflect reduced antigenic similarity. The $\log_2$ titre is modelled reflecting the two-fold serial dilution of antiserum in assays. We describe below how the variation in titres

attributable to the antigenic properties of viruses can be attributed to virus HA-gene sequences, firstly by mapping antigenic changes to branches of the phylogenetic tree and secondly by attributing antigenic changes to specific amino acid differences between reference and test viruses. Model selection was performed using BSSVS via binary indicator variables associated with each branch or amino acid position [18]. These indicator variables, also known as binary mask variables, take the value zero or one determining variable exclusion (masking) or inclusion, respectively. The optimal combination of branches, or amino acid positions at which substitutions explain antigenic changes, is therefore determined by sampling these binary mask variables using MCMC.

Throughout this approach, we model the assay titre Y measured for an antiserum raised against reference virus r and each specific virus v on a given date d as lognormally distributed:

$$log_2\ Y_{r,v,d} \sim Normal(H_{r,v} + D_d, \sigma_Y^2) \quad (1)$$

The $\log_2$ titre has a mean determined by combining the underlying $\log_2$ titre for each combination of reference virus and test virus, $H_{r,v}$, and an effect accounting for day-to-day experimental variability in titres, D_d. The use of $\log_2$ reflects the use of a two-fold serial dilution of antiserum, with recorded titres being the reciprocals of these dilutions. Residual variance of measured titres around this mean is represented by σ_Y^2. To improve the succinctness and therefore clarity of this section, prior distributions and other implementation details are provided in the *Model Implementation* subsection of the *Materials and Methods*, and details of indices, parameters, variances and other minor terms are described in Table 1.

The underlying $\log_2$ titre in **Eq 1** is itself normally distributed:

$$H_{r,v} \sim Normal(I_r + A_v - \Delta_{r,v}, \sigma_H^2) \quad (2)$$

Underlying titres are modelled as depending on three general characteristics of the assayed viruses and antisera. The contributions of effects for reference strain immunogenicity, I_r, test virus avidity, A_v, and the antigenic relationship, $\Delta_{r,v}$ are each inferred. The immunogenicity and avidity terms reflect general reactivity of antisera and viruses respectively. These manifest as a trend for higher or lower titres against all viruses or antisera for which titres are measured, independent of antigenic relationships. The genetic determinants of this antigenic component is of principal interest, so in the remainder of this section we describe how variation in this term is attributed to differences in the HA protein (see Table 1 for further details on the other effects). Since antigenic differences between reference and test viruses manifest as lower titres, the antigenic component of the model is constrained to take only non-negative values and is subtracted from the other terms in the model. $\Delta_{r,v}$ is defined in several different ways in Eqs 3, 4, 5 and 6 below, to allow us to compare different models of this antigenic relationship.

Antigenic difference is initially modelled as a linear combination of effects that occur during the phylogenetic evolution of the assayed viruses, with terms representing every branch, ψ, of the phylogeny to which amino acid substitutions were mapped in an ancestral state reconstruction (see Materials and Methods), Ψ, tested as predictors of reduced HI titres:

$$\Delta_{r,v} = \sum\nolimits_{\psi \in \Psi} \gamma_\psi m_\psi \delta_\psi(r, v) \quad (3)$$

For each branch, a precomputed indicator variable $\delta_\psi(r, v)$ is one when the branch falls on a direct path through the tree separating the reference and test viruses, and zero otherwise. Consequently, the titre for each combination of reference virus and test virus only depends on the antigenic weights for the combination of branches that fall between them on a path traced through the tree. The parameter m_ψ is the antigenic weight associated with the branch, the

Table 1. Model indices, terms and parameters.

Type	Term	Explanation
Indices	r	Reference virus used to generate ferret antisera
	v	Test virus evaluated as antigen in an assay
	d	Date on which experiment was carried out
	ψ	A branch
	λ	An amino acid position
	κ	A substitution at a specific amino acid position
Sets	Ψ	The set of branches of the HA phylogeny inferred to contain amino acid substitutions by phylogenetic analysis (see Materials and Methods)
	Λ	The set of non-conserved amino acid positions
	K_λ	The set of observed substitutions at a specific amino acid position
	$\hat{\Psi}$	$= \{\psi \in \Psi : \bar{\gamma}_\psi > p\}$ The subset of the branches in Ψ inferred from Eq 3, with probability $> p$, to map to antigenic changes
Indicator variables	$\delta_\psi(r, v)$	1 if a branch separates reference and test virus, 0 if not
	$\delta'_\kappa(r, v)$	1 if substitution separates reference and test virus, 0 if not
Data	$Y_{r,v,d}$	The measured titre for an antiserum raised against a particular reference virus and a test virus on a particular date
	$F_{1,\lambda}$, $F_{2,\lambda}$	Structural feature scores associated with an amino acid position (scaled to lie between zero and one)
Binary masks (determining whether a term is included in the model)	$\gamma_\psi(\bar{\gamma}_\psi)$	Binary mask determining whether to include (1) or mask (0) an antigenic term associated with a branch in the HA phylogeny (and its posterior mean)
	ζ_λ	Binary mask determining whether to include or mask an antigenic term associated with an amino acid position (in a structurally-naïve model).
	$\breve{\zeta}_\lambda$	Binary mask determining whether to include or mask an antigenic term associated with an amino acid position (in a structurally-aware model).
Antigenic and related terms	$H_{r,v}$	Underlying $\log_2$ titre for reference virus and test virus
	D_d	Experimental effect, for instance due to reagent variability or temperature, that results in a general tendency for higher or lower titres on a given day
	I_r	Reference virus immunogenicity effect that results in antisera with higher or lower titres
	A_v	Test virus avidity effect that results in a tendency for higher or lower titres, which is attributed to differences in avidity for the virus receptors on the red blood cell
	$\Delta_{r,v}$	The antigenic difference between reference strain and test virus
	m_ψ	Antigenic weight mapping to a specific branch
	m'_κ	Antigenic effect of a specific substitution at an amino acid position
	$\tilde{m}_\lambda$	Component of antigenic effect ascribed to an amino acid position
	$\tilde{m}'_\kappa$	Component of antigenic effect ascribed to a specific amino acid substitution
Probabilities	p	Threshold posterior mean probability for identifying a branch as potentially mapping to an antigenic change from Eq 3
	$\pi(\bar{\pi})$	Probability of branch being inferred to be antigenically important (and its posterior mean), prior for γ_ψ
	$\pi'(\bar{\pi}')$	Structurally-naïve probability of any amino acid position being inferred to be antigenically important (and its posterior mean), prior for ζ_λ
	$\pi_\lambda(\bar{\pi}_\lambda)$	Structure-informed probability of a specific amino acid position being inferred to be antigenically important (and its posterior mean), prior for $\breve{\zeta}_\lambda$
Powers	ρ_1, ρ_2	Power terms control importance of structural features in calculating π_λ
Variances	σ_Y^2	Residual variance of data around underlying titre and date effect
	σ_H^2	Variance of underlying titres around modelled effects

https://doi.org/10.1371/journal.pcbi.1010885.t001

expected drop in $\log_2$ HI titre when two viruses are separated in the phylogeny by branch ψ. The binary mask variable, γ_ψ, takes the value zero or one determining whether branch ψ is either excluded from or included in the model, respectively, and each antigenic effect, m_ψ, represents the antigenic effect of a specific branch when it is included ($\gamma_\psi = 1$). When γ_ψ is zero, any antigenic weight attributed to the branch is nullified (as the product, $\gamma_\psi m_\psi$, is zero). A higher proportion of MCMC samples with $\gamma_\psi = 1$ indicates higher support in the data for an antigenic change mapping to branch ψ. For each branch, the proportion of $\gamma_\psi = 1$, which is also the posterior mean value, $\bar{\gamma}_\psi$, is referred to as the inclusion probability for the branch. This allows antigenic changes in the evolution of assayed viruses to be mapped to specific branches of the phylogeny to generate $\hat{\Psi} = \{\psi \in \Psi : \bar{\gamma}_\psi > p\}$, the subset of those branches tested that are inferred to be potentially antigenically significant by having inclusion probability above some threshold, p. In previous work these phylogenetic variables were selected using a random restart hill-climbing algorithm that optimised Akaike Information Criterion (AIC) to reduce computation cost [12,13]. However, we have subsequently shown that variable selection using these binary masks is a superior strategy [18].

Incorporating amino acid substitutions. Next, terms were introduced to explicitly attribute antigenic differences between viruses to specific amino acid changes. The shared evolutionary history of the viruses has the potential to facilitate false statistical support for substitutions due to repeated measurements. To control for the evolutionary relationship between viruses and reduce this risk, $\hat{\Psi}$, the subset of phylogenetic terms identified as explaining antigenic variation using **Eq 3**, were retained in the new model. Branch variables were ranked by their inclusion probability $\bar{\gamma}_\psi$ and p, the threshold inclusion probability, was chosen so that the proportion of branches from Ψ that were carried forward in $\hat{\Psi}$ was $\bar{\pi}$, the posterior mean inclusion probability of a branch.

The subsequent model then included all of these terms (to control for the shared evolutionary history of the viruses) together with terms representing amino acid positions:

$$\Delta_{r,v} = \sum_{\psi \in \hat{\Psi}} m_\psi \delta_\psi(r,v) + \sum_{\lambda \in \Lambda} \zeta_\lambda \sum_{\kappa \in K_\lambda} m'_\kappa \delta'_\kappa(r,v) \tag{4}$$

Here, each amino acid position within the set of non-conserved positions Λ is indexed by λ, while all observed amino acid substitutions at each position λ are in turn indexed by κ. Each position was associated with a binary mask variable, ζ_λ, that takes the value zero or one, as above, determining whether substitutions at position λ contribute to variation in titres. A precomputed indicator variable, $\delta'_\kappa(r, v)$, indicates whether or not each specific amino acid difference separates the reference and test viruses. Each antigenic effect, m'_κ, represents the antigenic effect of a specific substitution.

Using the model described in **Eq 4**, antigenic effects of alternative substitutions at the same amino acid position are independent. An alternative model where the antigenic effects of different substitutions at the same position are linked is also explored:

$$\Delta_{r,v} = \sum_{\psi \in \hat{\Psi}} m_\psi \delta_\psi(r,v) + \sum_{\lambda \in \Lambda} \zeta_\lambda \tilde{m}_\lambda \sum_{\kappa \in K_\lambda} \tilde{m}'_\kappa \delta'_\kappa(r,v) \tag{5}$$

The key difference here from **Eq 4** is that the antigenic effect of a substitution is partitioned into a position-specific component, $\tilde{m}_\lambda$, and a substitution-specific component, $\tilde{m}\prime_\kappa$. The position-specific component is shared by every substitution observed at an amino acid position. This reflects an expectation that a range of substitutions at more antigenically important positions will tend to have greater antigenic impacts than at other positions. The substitution-

specific component, on the other hand, allows for variability in the antigenic impact of alternative substitutions at the same position.

Incorporating structural information. Eqs 4 and 5, above, describe how variation in titres was attributed to antigenic effects of amino acid substitutions in what we term 'structurally-naïve' models. We now incorporate information from analysis of 3-D protein structure into the above models by using it to inform the prior probability that substitutions at an amino acid position are involved in antigenic evolution (by influencing the binary mask variable associated with each position). We implement this approach using two structural features: proximity to the receptor-binding site (RBS), and a predicted epitope score derived from a tool used to predict conformational epitopes from tertiary protein structure (see Materials and Methods). However, this approach is not limited to the use of these features and could be adapted to either a single or more than two structural features. Here, we refer to the two structural features as F_1 and F_2.

In Eqs 4 and 5 (the structurally-naïve models), the binary mask terms associated with each amino acid position, ζ_λ, share a common prior distribution and therefore are equally likely to be included in the model prior to observation of the data. Here, the structurally-aware version of the models described above retain the structure described in Eqs 4 and 5. The only difference is that the binary mask variable $\check{\zeta}_\lambda$ is redefined so that the probability of each position being selected may be influenced by structural features $F_{1,\lambda}$ and $F_{2,\lambda}$:

$$\Delta_{r,v} = \sum_{\psi \in \dot{\Psi}} m_\psi \delta_\psi(r,v) + \sum_{\lambda \in \Lambda} \check{\zeta}_\lambda \tilde{m}_\lambda \sum_{\kappa \in K_\lambda} \tilde{m}'_\kappa \delta'_\kappa(r,v) \tag{6}$$

$$\check{\zeta}_\lambda \sim Bernoulli(\pi_\lambda) \tag{7}$$

$$\pi_\lambda = F_{1,\lambda}^{\rho_1} \cdot F_{2,\lambda}^{\rho_2} \tag{8}$$

For each position, the outcome of Bernoulli trial which determines the value of the binary mask variable now depends on both the data and a position-specific, structure-informed prior probability of antigenic importance π_λ. As described in **Eq 8**, the probability term π_λ is computed directly from the structural features associated with position λ (which are scaled between zero and one) and the power terms ρ_1 and ρ_2, the values of which are fitted to the data. A tendency across all positions for those with higher values for the structural features to be involved in antigenic change will result in higher estimates for ρ_1 and ρ_2. Conversely, if the data do not support a relationship between a structural feature and antigenicity, the associated ρ term will tend towards zero.

Results

To model the genetic basis of antigenic differences between influenza viruses, we first compiled antigenic (HI) and HA genetic sequence data from A(H3N2) viruses isolated during the period 1990–2014, and used the sequence data to construct a phylogeny describing the evolutionary relationships between viruses. Next, we mapped variation in HI titres to branches of the phylogenetic tree–BSSVS was used to identify branches of the phylogeny representing amino acid substitutions causing antigenic change and to quantify the associated degree of antigenic change. Specifically, we calculated the proportion of MCMC samples in which a phylogenetic variable from **Eq 3** is selected (where the associated binary mask variable, $\gamma_\psi = 1$), which represents confidence in the selected variable, and we also recorded a quantitative estimate of the antigenic change, m_ψ (in $\log_2$ HI units), associated with the branch.

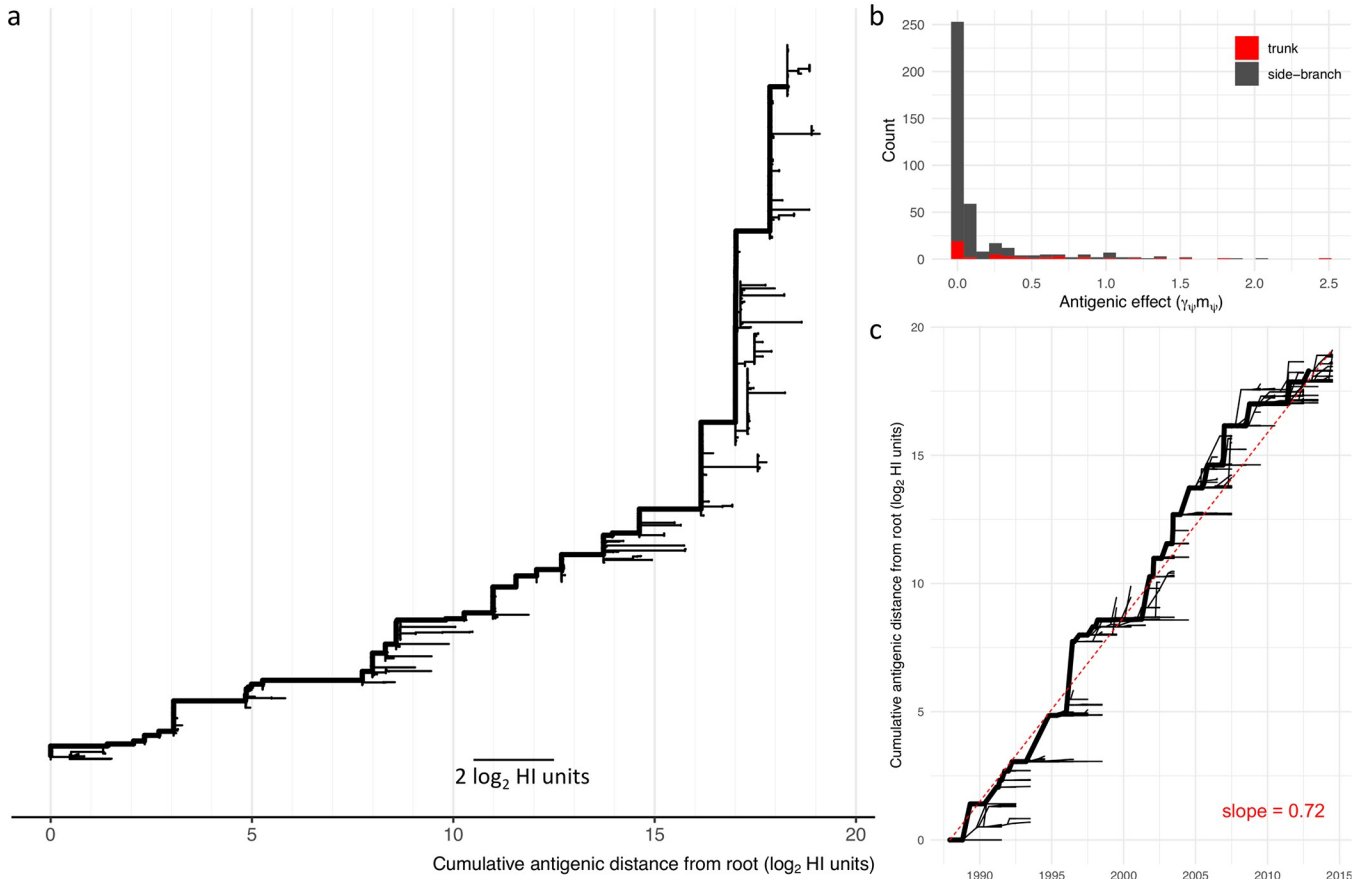

Fig 1. Antigenic evolution mapped to HA phylogeny. Antigenic change, as expressed in HI titres, mapped to branches of the HA1 phylogeny. The antigenic effect ($\log_2$ HI units) for each branch estimated as the average drop in titre when virus and antiserum separated by the branch are tested together. (**a**) HA phylogeny with branch lengths scaled to show antigenic effects ($\gamma_\psi m_\psi$). The x-axis shows cumulative antigenic distance from the root. The trunk lineage is shown as a thick line. (**b**) Histogram showing antigenic effects ($\gamma_\psi m_\psi$) estimated for each branch in the trunk lineage (red) and in the side branches (grey). (**c**) Phylogeny is plotted to show time on the x-axis (years) and cumulative antigenic distance from the root on the y-axis ($\log_2$ HI units). The trunk lineage is shown as a thick line. Dashed red line indicates the linear regression between time since the root and antigenic distance from the root for each node in the phylogeny (slope = 0.72).

https://doi.org/10.1371/journal.pcbi.1010885.g001

The contributions of each branch to antigenic evolution are shown in the phylogenetic tree in Fig 1a where branch lengths indicate the posterior mean value of $\gamma_\psi m_\psi$ in Eq 3. In this visualisation, the horizontal dimension expresses antigenic change. Consequently, an antigenically homogenous clade will appear as a vertical line, regardless of the amount of molecular evolution that has occurred within it or the time spanned. The mean number of branches included in the model in an individual step of the MCMC was 64 (95% HPD, 55–75) out of 397. Branches in the trunk lineage were more likely to be included in the model explaining variation in HI titres, compared with branches in the rest of the tree (odds-ratio 5.0; 95% CI, 3.4–7.4). The histogram in Fig 1b illustrates a highly right-skewed distribution of antigenic weights assigned to branches with relatively few antigenic events of more than 1 $\log_2$ HI units. While the histogram shows branches with larger effects to be found in both the trunk and side branches, the rate of antigenic change was found to be considerably higher in the trunk. Rates of antigenic evolution in the trunk and side branches were calculated by summing antigenic weights ($\gamma_\psi m_\psi$) and dividing by the sum of branch lengths measured in years of evolutionary time estimated using a molecular clock analysis implemented in BEAST [20,21]. The rate of antigenic drift in the trunk lineage was estimated to be 0.73 $\log_2$ HI units per year (95% HPD,

0.67–0.78), compared with 0.05 $\log_2$ HI units per year in the rest of the tree (95% HPD, 0.04–0.06). Cumulative antigenic distance from the root was calculated across the phylogenetic tree by summing antigenic effects ($\gamma_\psi m_\psi$) across each branch in a path between the root and every internal node and tip. This cumulative antigenic distance is represented in the horizontal dimension in the tree in Fig 1a. Linear regression of cumulative antigenic distance for each node and cumulative branch lengths estimated a rate of antigenic drift of 0.72 $\log_2$ HI units per year (Fig 1c). This aligns very closely with a figure of 0.71 previously estimated for a dataset of A(H3N2) viruses from the period 1985–2015 using a comparable approach [8].

Genetic determinants of antigenic change

Next, amino acid substitutions were tested as predictors of reduced HI titres. Antigenic weights were estimated for each substitution with binary mask variables (ζ_λ) dictating whether substitutions at a particular position were included in the genetic model of variation in titres. Based on the HI data, for each variable position the posterior mean value of ζ_λ, or the posterior inclusion probability, is the inferred probability that substitutions at position λ have contributed to antigenic evolution. To assess how well each model did in terms of attributing antigenic changes to amino acid substitutions, we examined posterior inclusion probabilities associated with amino acid positions in the context of antigenic sites defined in the literature. While the important antigenic areas of HA are known, definitions of the constituents and boundaries of the antigenic sites vary, making a binary in-or-out classification to assess model selection problematic. For this reason, we considered the distance in 3-D space from each residue to these sites rather than a binary in-or-out classification. These distances were calculated between alpha carbons, therefore were relatively insensitive to the changes in protein structure occurring during evolution. For example, distances calculated using the HA structures of A/Aichi/2/68 and A/Brisbane/10/2007 were highly correlated (R^2 >0.99, S1a Fig), despite the two viruses being separated by 39 years of evolution. At each step of the MCMC, the mean distance to an antigenic site, averaged across the set of positions selected by the model ($\zeta_\lambda = 1$) at that step, was calculated. The mean distance to antigenic site averaged over the MCMC chain was used to evaluate model performance with lower values indicating higher ability to correctly attribute antigenic variation to causative substitutions. Initial comparison of structurally-naïve models showed that a model assuming a link between the antigenic impact of substitutions occurring at the same HA position (antigenic relationships, $\Delta_{r,v}$, modelled using Eq 5) outperformed a simpler model that assumed no such link ($\Delta_{r,v}$ modelled using Eq 4) as indicated by a lower mean distance to antigenic sites of amino acid residues inferred to have substitutions explaining antigenic differences (Table 2). Using a single effect size (m'_κ in Eq 4 or $\tilde{m}'_\kappa$ in Eq 5) for forward and reverse substitutions (symmetric substitution effects) also led

Table 2. Model performance evaluated by distance from antigenic sites of positions implicated in antigenic evolution, averaged across posterior samples.

Model	Symmetric substitution effects	Mean distance to antigenic site (Å)
Structurally-naïve model—substitution model with independent effect sizes (Eq 4)	Yes	11.80
	No	13.28
Structurally-naïve model—substitution model with effect sizes linked to position (Eq 5)	Yes	6.86
	No	11.39
Structurally-aware model–structure influencing inclusion probabilities for positions and effect sizes for substitutions linked to position (Eq 6)	Yes	2.55
	No	3.05

https://doi.org/10.1371/journal.pcbi.1010885.t002

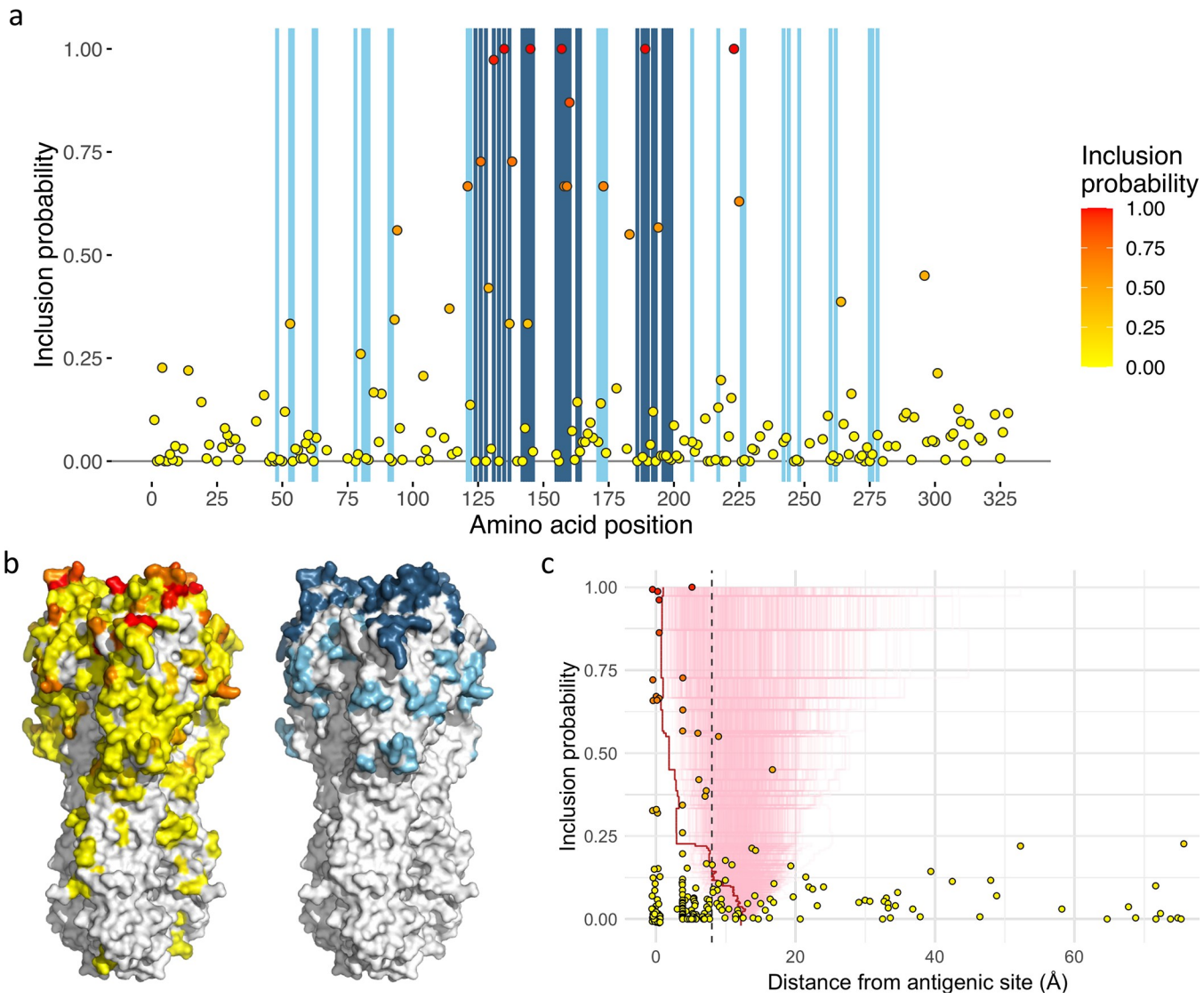

Fig 2. The role of HA positions in antigenic evolution estimated using a structurally naïve model. (**a**) Each point represents the posterior inclusion probability for a variable HA1 position. Blue vertical shading indicates positions on the x-axis that are described in antigenic sites A and B (dark blue) and antigenic sites C, D, and E (light blue). (**b**) Amino acid positions on the surface model of a HA structure are coloured: to the left, by inclusion probability following the colour scheme in **a**; to the right, antigenic sites A and B are shown in dark blue and sites C, D and E in light blue. (**c**) Posterior inclusion probability for each variable amino acid position is plotted against distance from the closest residue in a defined antigenic site. Points at 0Å on the x-axis, representing residues in described antigenic sites, have been adjusted in both dimensions to allow visualisation of overlapping points. A red line indicates the mean distance of residues selected with inclusion probability equal to or higher than that on the y-axis. Light red lines show a null distribution derived from 1,000 randomisations retaining the inclusion probabilities from the model fitted to the data. A dashed vertical line at 8Å marks a threshold at which a position is approximately two residues away from an amino acid in a described antigenic site.

https://doi.org/10.1371/journal.pcbi.1010885.g002

to better model performance compared with estimating two effect sizes (asymmetric substitution).

For the best performing structurally-naïve model (described by **Eq 5**), inclusion probabilities for each position with substitutions present in the dataset are indicated in **Fig 2a**, with vertical bars showing the locations of antigenic sites [22]. Data underlying Fig 2 are available in S1 Data. In **Fig 2b**, residues on a structural model of HA are shown coloured by inclusion probabilities and the locations of antigenic sites are shown to the right. Six HA positions were

identified with very high confidence, each being associated with an inclusion probability of at least 0.95: positions 131, 135, 145 belonging to antigenic site A, 157 and 189 to antigenic site B, and 223 located on an exposed loop on the boundary of the RBS. A posterior mean value of 0.12 for the probability of inclusion, $\bar{\pi}'$, was used to determine the inclusion probability required for inclusion of an HA position in an optimal model. Ranking inclusion probabilities and taking the top 12th percentile corresponded to 25 HA positions. Of the top 25 positions ranked by inclusion probability, six belonged to antigenic site A (126, 131, 135, 137, 144 and 145), five to site B (157, 158, 159, 160 and 189), one to site C (53), two to site D (121 and 173) and none to site E. In **Fig 2c**, each HA position tested is positioned by its distance from the closest antigenic site and its posterior inclusion probability. When ranked by inclusion probability, the top 16 positions fall within 8Å of the core antigenic site limits, and only two positions above this threshold (183 at 9.0Å and 296 at 16.7Å) were included (with a probability greater than the threshold of 0.33). However, **Fig 2c** shows substitutions at some more distant positions were being included in the model in a non-negligible proportion of MCMC samples (positions 4 at 75.7Å and 14 at 52.3Å were associated with inclusion probabilities of 0.23 and 0.22 respectively). Both of these positions are in the stalk domain distant from the antigenic sites around the RBS and therefore it is deemed very unlikely that substitutions at these positions have contributed to antigenic evolution as assessed in HI assays. The occasional inclusion of parameters associated with positions such as these inevitably disturbs estimates for other parameters within the model.

Incorporating structural data

To investigate whether incorporating data on protein structure could assist the accuracy with which variables explaining antigenic differences could be identified, the locations of residues within the HA structure were used to influence the inclusion probabilities estimated for each position. Two structural features were considered, the distance from the RBS and a structure-based epitope score which estimates how accessible areas of a protein are to an antibody. For each residue, its distance to the RBS was calculated as minimum distance in 3-D space of its alpha carbon to the closest alpha carbon of a RBS residue. A structure-based epitope score was calculated for each HA residue using BEpro [23] and reflects how exposed and accessible for antibody binding, the region centred on each residue is. Comparison of these two measures calculated across solved HA structures from three A(H3N2) viruses isolated during the period covered by the HI dataset (A/Finland/486/2004, A/Hong Kong/4443/2005 and A/Brisbane/10/2007) and an earlier, evolutionary founder virus from the 1968 pandemic (A/Aichi/2/68) indicated that changes to protein structure during virus evolution did not greatly affect these measures. The variances in measurements made between different HA structures were small indicating that structural information used is not unduly influenced by the choice of HA structure. For example, comparison of RBS distances and epitope scores for the HAs of A/Aichi/2/68 and A/Brisbane/10/2007 were highly correlated (R^2 >0.99 and 0.94 respectively, **S1a and S1c Fig**). Therefore, mean distance from the RBS and epitope scores averaged over the four structures (**Fig 3**) were considered suitable for use in modelling. Data underlying Fig 3 are available in S2 Data.

The structural measurements plotted in **Fig 3** were allowed to influence the selection of amino acid substitutions inferred to explain variation in the antigenic component of HI titres, $\Delta_{r,v}$, according to **Eqs 6–8.** Structural information was used to estimate a structure-informed probability (π_λ) for each HA position, with the parameters ρ_1 and ρ_2 determining the importance of the proximity to the RBS (denoted $F_{1,\lambda}$) and the structure-based epitope score (denoted $F_{2,\lambda}$) of each position, according to **Eq 8**. In **Fig 4a** each HA residue is positioned in a

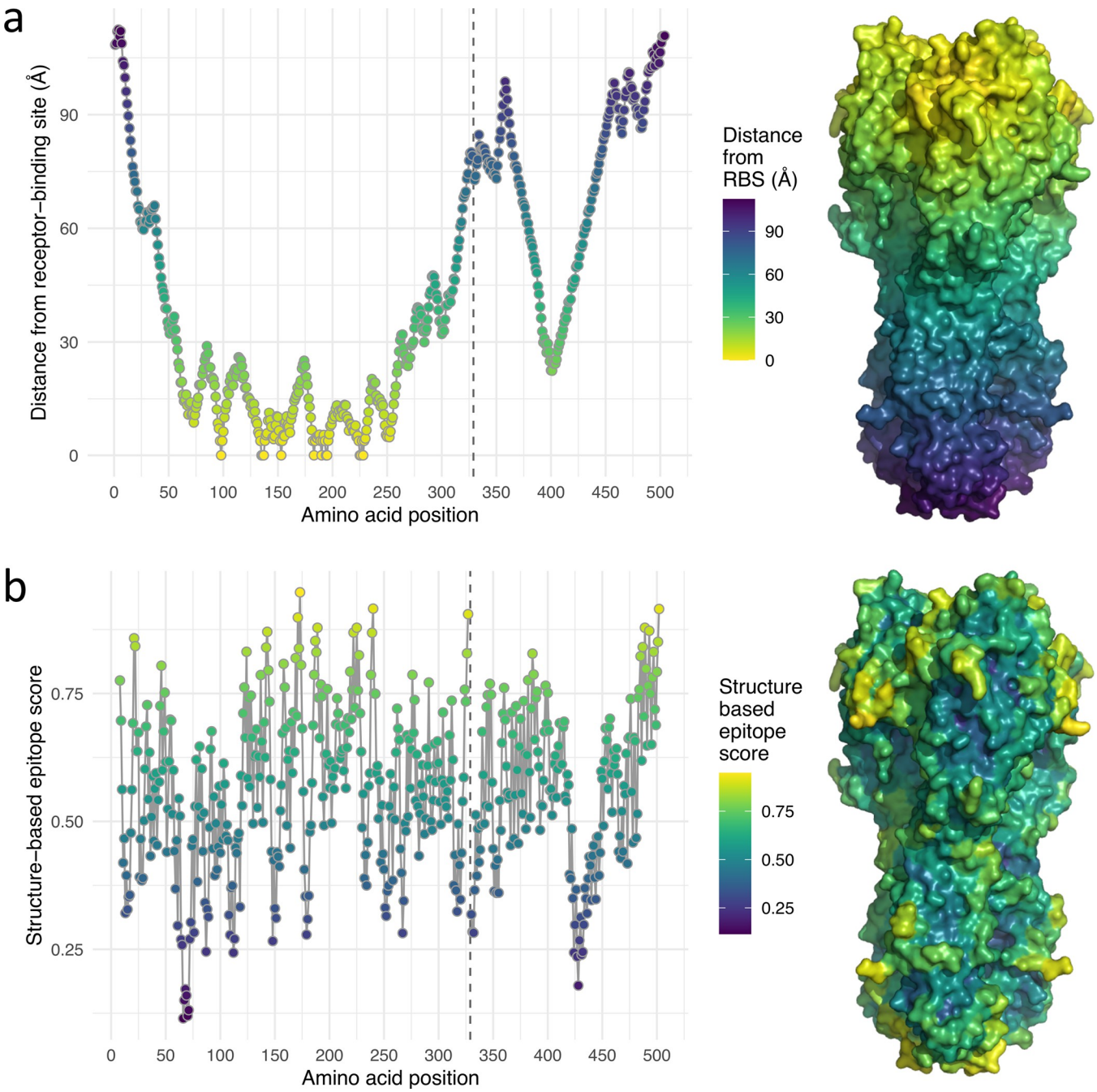

Fig 3. Structural features of the influenza A(H3N2) HA. (**a**) The distance of each HA residue to the closest of the residues comprising the RBS. To the right, a surface representation of the HA structure is shown coloured according to the distance key. (**b**) The structure-based epitope score for each HA residue was calculated using BEpro [23]. To the right, a surface representation of the HA structure is shown coloured according to the epitope score key. In each plot, a vertical dashed line at position 329 indicates the boundary between HA1 and HA2.

https://doi.org/10.1371/journal.pcbi.1010885.g003

2-D space according to these two measures, while the colouring shows the structure informed probability of antigenic importance calculated from posterior mean values for the ρ parameters according to **Eq 8**. Data underlying Fig 4 are available in S3 Data. **Fig 4a** shows that neither the proximity to the RBS nor the structure-based epitope score fully dominated the determination of the probability, however **Fig 4b** shows the ρ parameter associated with RBS proximity

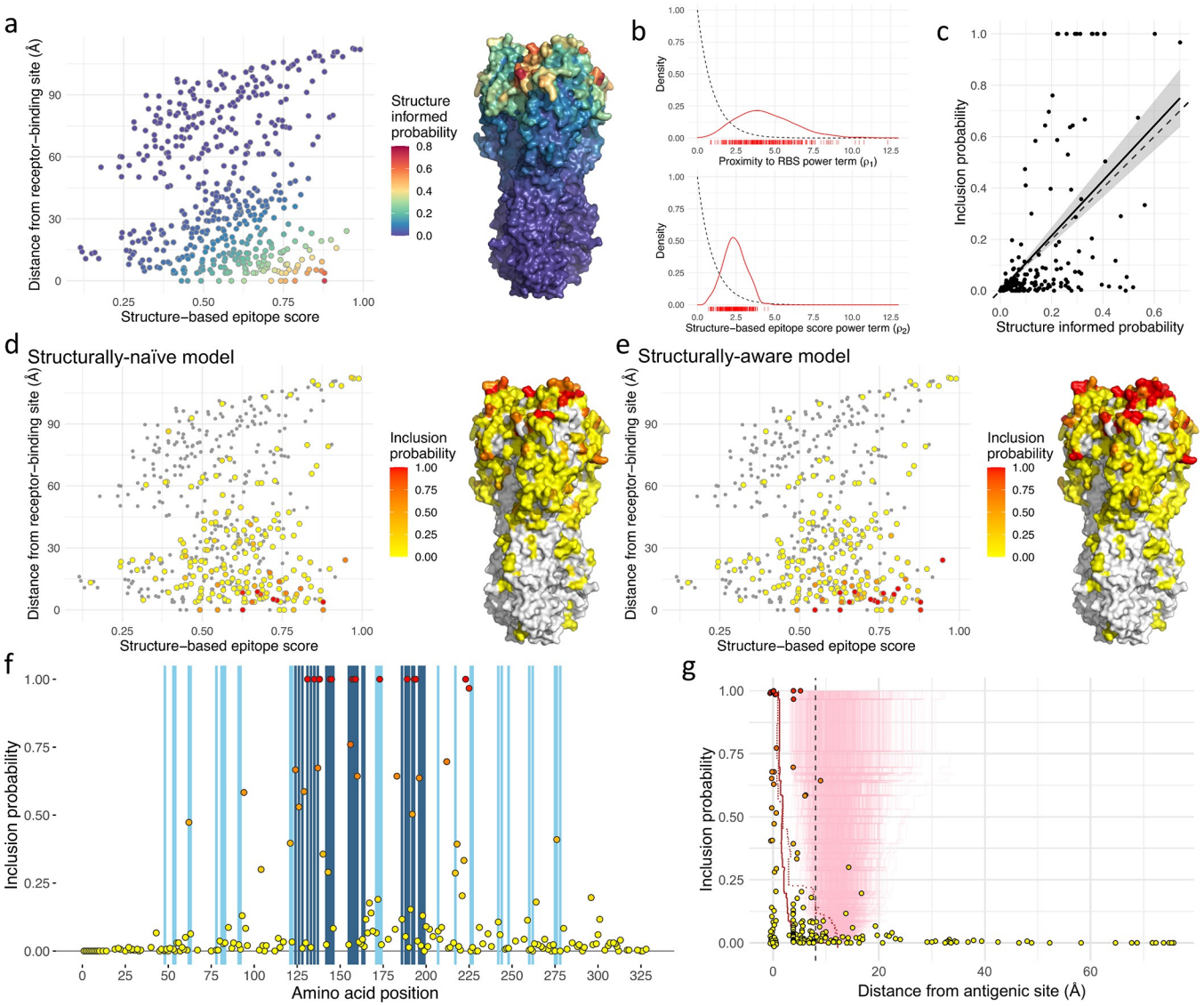

Fig 4. Use of structural data to guide variable selection. Plots summarise variable selection using a model where posterior inclusion probability that substitutions at an HA position affect antigenicity is impacted by a structure-informed probability that depends on a structure-based epitope score and distance from the RBS (**a**). HA residues are positioned according to a structure-based epitope score and distance from the RBS. Colour indicates the structure-informed probability of antigenic importance. To the right, structure informed probability is shown on the HA structure. (**b**) Posterior distributions for power terms that link proximity to the RBS (ρ_1, top) and structure-based epitope score (ρ_2, bottom) for each HA position to a structure-informed probability for the position, π_λ, according to Eq 8. Individual values sampled from the posterior distribution are shown below the x-axis. Prior distributions for these parameters, defined as Gamma(1, 1), are shown as dashed black lines. (**c**) Scatterplot showing the relationship between the structure-informed probability and the posterior inclusion probability. The solid black line has a slope of 1.07 alongside a dashed line of slope 1, with standard error from the linear model indicated in grey. (**d-e**) HA residues are positioned according to structure-based epitope score and distance from the RBS, identical to **a**. The colour scheme indicates posterior inclusion probabilities estimated using structurally-naïve (**d**) and structurally-aware (**e**) models. Positions without substitutions are shown in grey. To the right of each scatterplot, inclusion probabilities are shown on the HA structure. (**f**) Each point represents the posterior inclusion probability for a variable HA1 position. Blue vertical shading indicates positions on the x-axis that are described in antigenic sites A and B (dark blue) and antigenic sites C, D, and E (light blue). (**g**) Inclusion probability for each variable amino acid position is plotted against distance from the closest residue in a described antigenic site. Points at 0 Å on the x-axis representing residues in described antigenic sites, have been adjusted in both dimensions to aid visualisation of overlapping points. A red line indicates the mean distance of residues selected with inclusion probability equal to or higher than that on the y-axis. Light red lines show a null distribution derived from 1,000 randomisations retaining the inclusion probabilities from the model fitted to the data. A dotted red line shows the corresponding line from a structurally-naïve model shown in Fig 2c. A dashed vertical line at 8Å marks a threshold at which a position is approximately two residues away from an amino acid in the described antigenic site.

https://doi.org/10.1371/journal.pcbi.1010885.g004

(mean, 4.43; 95% HPD, 1.60–8.61) was higher than that associated with the epitope score (mean, 2.40; 95% HPD, 0.93–3.73), indicating proximity to the RBS to be particularly useful as a predictor, given both structure features were rescaled between zero and one for modelling. As expected, the structure-based probability does not fully determine whether or not a position is included in the model but guides variable selection, according to Eq 7, when correlations between patterns of substitution obscure relationships between antigenic change and causative substitutions. The correlation between the structure-based probability, $\bar{\pi}_\lambda$, and the posterior inclusion probability, $\bar{\zeta}_\lambda$, has a slope of 1.07 as shown in Fig 4c. Variation around the slope is expected as other factors such as the side-chain properties of the actual amino acid substitutions must contribute to determine whether positions influence antigenicity, however the slope of close to 1 indicates that the relationship between structure-informed probability and posterior inclusion probability is fitted correctly. In Fig 4d and 4e, HA positions are again placed according to structure-based epitope scores and distance from the RBS. Each non-conserved HA position is coloured by its posterior inclusion probability, the confidence that substitutions at the position have impacted antigenic evolution, in models fitted without (Fig 4d) and with structural data (Fig 4e). The effect of structural information guiding variable selection is clear with residues positioned in the bottom right corner of the scatterplot for the structurally-aware model tending to be associated with higher inclusion probabilities, which corresponds to a greater number of red residues in surface-exposed areas near to the RBS in Fig 4e compared with Fig 4d. Notably, a higher proportion of positions with an inclusion probability of one are identified using the structurally-aware model (Fig 4f compared with Fig 2a).

Incorporating structural information into variable selection reduced uncertainty in model identification. Using the structurally-aware model, 144 of 199 (72.4%) variable HA positions were either included in the model or excluded in at least 95% of MCMC samples (inclusion probability >0.95 for 14 positions and <0.05 for 130, S1 Table), while for the structurally-naïve model the corresponding number was only 119 (59.8%) (>0.95 for 6 positions and <0.05 for 113, S1 Table). In Fig 4f, posterior probabilities are shown for each HA position where substitutions were present with blue shading used to indicate antigenic sites. This plot shows that the majority of selected positions either belonged to defined antigenic sites or were very close to one of them in primary amino acid sequence. Of the 14 positions associated with an inclusion probability of at least 0.95, four were in antigenic site A (positions 131, 135, 144 and 145), five in site B (157, 158, 159, 189 and 193) and one in site D (173), while the others were either defined as belonging to the RBS (194 and 225) or were located close to the RBS (138 and 223). Incorporating structure into variable selection resulted in greater accuracy as quantified by the distances of amino acid residues from defined antigenic sites (Fig 4g and Table 2), with a higher effective number of parameters contributing to antigenic distance (S1 Table). Each of the 18 residues with the highest inclusion probability belonged to defined antigenic sites or were within 8Å of these and only position 183, at 8.9Å, was further away and associated with an inclusion probability of >0.5. Of all positions associated with an inclusion probability of >0.05, position 43, 19.3Å away, with an inclusion probability of 0.08, was the most distant. For comparison, 19 positions further than 19.3Å from antigenic sites were associated with inclusion probabilities of at least 0.05 using the structurally-naïve model (Fig 2c), up to a maximum of 75.7Å (position 4, inclusion probability of 0.23).

Previous work by Koel *et al.* examined the genetic basis of the largest antigenic changes through systematic testing by reverse genetics of amino acid differences mapping to transitions between clusters apparent on antigenic maps generated from HI data for viruses spanning 1968–2003, identifying a key role for substitutions at HA positions 145, 155, 156, 158, 189 and 193 and to a somewhat lesser extent positions 133 and 135 [24]. We identify substituions in

common with that approach such as 156QH acquired by A/Fujian/411/2002-like viruses and observe that substitutions at these 'Koel' sites continued to play an important role in antigenic drift in the period beyond 2003. For example, in the following period several substitutions at these positions appear in trunk lineage branches to which significant antigenic drift was attributed: 159YF, 145KN, 193SF, 189NK, and 145NS. The highest average antigenic weight attributed to a substitution, averaged across titres, in the studied dataset was for amino acid substitution 189NK (1.73 $\log_2$ HI units). Previous work has shown substitutions 189QK and 189KR contributed to changes between A(H3N2) antigenic clusters in the 1970s and 1980s [24]. In the period studied, the substitution 189NK appears in the trunk lineage in one of the most antigenically significant branches separating A/Perth/16/2009-like viruses from earlier A/California/7/2004-like viruses. Interestingly the substitution 189NK also appeared in a side branch at a similar point in time leading to an evolutionarily unsuccessful lineage. In the trunk lineage, 189NK co-occurred with 158KN, which itself was present in multiple branches and has an estimated mean antigenic impact of 0.94 $\log_2$ HI units and 212TA which also mapped to two terminal branches though estimated to have a lower antigenic impact (0.29 $\log_2$ HI units). Substitutions co-occurring with 189NK in the side branch are examined below. Excepting 189NK, the next highest antigenic weight assigned to a substitution was for 135KT (1.65 $\log_2$ HI units). This substitution became fixed in the A(H3N2) population in the mid-1990s and was quickly followed by the incorporation of 133DN into the trunk lineage, a substitution combination that introduced an additional *N*-linked glycosylation motif, NGT, across positions 133–135.

The highest antigenic effect (1.69 $\log_2$ HI units) for an amino acid substitution outside of the 'Koel' positions was between A and T at position 131, which falls in antigenic site A. The substitution 131AT was incorporated into a branch forming the trunk lineage with 131T viruses rising in frequency to achieve fixation over the period ~2002–2004. After position 131, the highest antigenic weights estimated at 'non-Koel' sites were at HA positions 144 and 194, which both neighbour 'Koel' sites. Position 144 is of note as multiple changes across the phylogeny away from 144N result in antigenic effects in concert with a loss of a potential *N*-linked glycosylation site (as 146S is almost entirely conserved among A(H3N2) viruses during the period studied). Indeed, each of the two highest antigenic effects mapping to internal side branches leading to more than only 1–3 viruses included loss of a glycosylation motif at positions 144–146 with 142RG and 144ND mapping to one and 62EK, 144NK, 158KN, and 189NK mapping to the other, with estimated antigenic weights of 2.48 and 1.41 $\log_2$ HI units respectively.

Bayesian model averaging delivers accurate prediction of HI titres

The full dataset was comprised of 38,757 titres which included 3,477 different virus and reference antiserum combinations. These combinations were measured with 1,737 different viruses (including reference viruses) and antisera raised against 151 reference viruses. To assess the predictive performance of genetic models of antigenic phenotype, a range of model variants were tested for their capacity to recover HI titres under two prediction schemes. Firstly, to assess the capacity of models to predict unobserved antigenic relationships, 10% of virus and reference antiserum combinations (348 combinations), were randomly selected 100 times, and all titres associated with those combinations were removed to act as test data. Models were trained using the remaining 90% of virus and reference antiserum combinations and tested for their ability to accurately predict the removed titres. Secondly, to evaluate model capacity to predict titres of uncharacterised viruses, 10% of viruses (174 viruses) were randomly selected 100 times, and all titres for these viruses measured using any antiserum were removed to act as

Table 3. Accuracy of genetic models in predicting antigenic phenotype. Measures of the difference between predicted and observed HI titres and the percentage of errors within 1 or 2 $\log_2$ HI units.

Model (Equation)	Symmetric substitution effects	Prediction scheme					
		1. Test-reference virus pairs			2. Test viruses		
		MAE	RMSD	% <1 (<2)[1]	MAE	RMSD	% <1 (<2)[1]
Phylogenetic model (Eq 3)	n/a	0.60	0.83	83 (94)	1.03	1.14	64 (87)
Structurally-naïve model—substitution model with independent effect sizes (Eq 4)	Yes	0.53	0.69	86 (99)	0.80	1.04	69 (94)
	No	0.53	0.69	86 (99)	0.81	1.05	69 (93)
Structurally-naïve model—substitution model with effect sizes linked to position (Eq 5)	Yes	0.52	0.69	87 (99)	0.80	1.04	69 (94)
	No	0.52	0.69	87 (99)	0.79	1.03	70 (94)
Structurally-aware model–structure influencing inclusion probabilities for positions and effect sizes for substitutions linked to position (Eq 6)	Yes	0.52	0.69	87 (99)	0.77	1.01	71 (94)
	No	0.52	0.69	87 (99)	0.77	1.02	71 (94)

MAE, mean absolute error; RMSD, root-mean-square deviation; n/a = not applicable. [1]The percentage of predicted titres within 1 (or 2) $\log_2$ HI units of the true underlying titre.

https://doi.org/10.1371/journal.pcbi.1010885.t003

test data with data for the remaining 90% of viruses used to train models. Under both prediction schemes, the models (**Eqs 3–6**) included indicator variables to determine whether a genetic variable contributed to differences in titres, in common with the previous sections. As these indicator variables are present, the combination of HA positions at which substitutions contribute to predictions may vary at each step of the MCMC. This therefore constitutes prediction by model averaging rather than prediction conditioned on a single best model, accounting for uncertainty in the identification of the substitutions that cause antigenic differences.

Employing the first scheme, titres were predicted initially with the antigenic component of titres, $\Delta_{r,v}$, modelled using the phylogenetic model described by Eq 3, with information on the antigenic changes associated with branches of the phylogeny ($\gamma_\psi m_\psi$) estimated using the training data only. This model provides a reasonable approximation of antigenic relationships, allowing titres to be predicted for unknown virus and reference strain combinations under cross-validation with a mean absolute error (MAE) of 0.60 $\log_2$ HI units (root-mean-square deviation RMSD = 0.83, Table 3, scheme 1). However, as the antigenic weights ($\gamma_\psi m_\psi$) associated with branches are purely additive, the tree model fails to account for the antigenic consequences of phenomena such as reverse substitutions or the presence of the same substitution in multiple branches.

To account for these non-additive antigenic events requires terms that explicitly describe the presence or absence of amino acid substitutions ($\Delta_{r,v}$ modelled using **Eqs 4, 5** or **6**). Using a model with independently estimated effect sizes for every substitution at each position (Eq 4) resulted in more accurate predictions with a MAE of 0.53 $\log_2$ HI units and RMSD of 0.69 (Table 3, scheme 1). Having the effects of substitutions at the same position linked (Eq 5) resulted in a marginal improvement in predictions (MAE = 0.52 and RMSD = 0.68) (Table 3, scheme 1). Improved predictions are expected under these substitution models, as they can account for convergent substitutions and reversions of substitutions. Using the structurally-aware model of antigenic relationships (Eq 6) resulted in similar prediction accuracy when predicting missing test-reference virus combinations (MAE = 0.52 and RMSD = 0.69, Table 3, scheme 1). The accuracy of these predictions are comparable to existing models for prediction of A(H3N2) HI titres [8]. This indicates that our efforts to increase the stringency of variable selection, rather than prioritising the selection of the maximally predictive set of genetic terms,

do not hamper the predictive ability of our approach. Under the best performing genetic model, 87% of predictions were made within 1 $\log_2$ HI unit of the observed titre and 99% within 2 $\log_2$ HI units (Table 3).

The accuracy of all models was reduced under this second prediction scheme (Table 3, scheme 2) with a MAE of only 1.13 $\log_2$ HI units (RMSD = 1.14) using the phylogenetic model. Predicting titres for viruses that are entirely absent from the training data is more challenging, in part, as it is not possible to estimate a virus avidity parameter (A_v in Eq 2), as recognised previously [8,12]. The accuracy of predictions was improved using the substitution-based models (MAE 0.77–0.81 $\log_2$ HI units and RMSD = 1.01–1.05). Interestingly, the structurally-aware model is slightly better performing here when predicting for cross-validation test datasets consisting of missing test viruses (0.77 compared with 0.79, Table 3, scheme 2). This indicates that more accurately attributing antigenic variation to the correct substitutions also offers an advantage when predicting antigenic relationships from HA sequences for viruses with no associated antigenic data.

In Fig 5a, titres predicted using scheme 1 are plotted against observed titres for the model with structural information (Eq 6). In Fig 5b, titres predicted using scheme 2 are plotted against the observed titres using the same model. Fig 5a shows a close relationship between predicted and observed titres while Fig 5b shows that in the absence of any information on the reactivity of a virus with any available antisera, there is a trend towards under-estimation of high observed titres. Such high titres tend to be associated with test viruses having higher than average titres across panels of antisera against which they are tested. Without information on these viruses (scheme 2), we observe a mean underestimation (predicted *v* observed) of -0.16 (Fig 5b), which compares with a corresponding value of -0.06 for scheme 1 (Fig 5a).

It should be noted that, in both the prediction schemes described above, there is non-independence of data points due to phylogenetic structure and the presence of viruses with highly similar evolutionary histories. To explore prediction in a related but independent dataset, we examined the predictive power of these models applied to VN measurements made between viruses collected in the years 2015–2020. In this forward prediction scheme, test datasets consisted of all titres for viruses collected in a specific year measured against antisera present in VN training data comprising viruses collected in previous years and associated antisera (Fig 5c). The null model was trained to the values of homologous titres for each reference virus in the training data only. The structural model not exposed to antigenic data was also trained using the values of homologous titres for antisera in the training data and additionally information from the structurally-aware model applied to the HI dataset (1990–2014): 1) the position-specific structure-informed inclusion probability (π_λ in Eq 7); and 2) the position-specific antigenic coefficient ($\tilde{m}_\lambda$ in Eq 6). Notably, this model trained to antigenic information in the HI dataset (1990–2014) is able to make informative predictions of antigenic relationships for viruses that evolved several years later, correlating with titres measured using a different assay (green line, Fig 5c). This prediction scheme mimics the application of the model and information from a data-rich situation to a data-poor situation. This indicates consistency in the characteristics of antigenically important sites and a degree of repeatability in the positions where substitutions affect titres. As expected, the model trained to information on the antigenic impact of substitutions on VN titres in the training datasets generally performs better (blue line, Fig 5c).

Discussion

Here, we describe an approach for identifying genetic changes that explain the antigenic drift of a rapidly evolving virus pathogen. The evolutionary process results in highly correlated

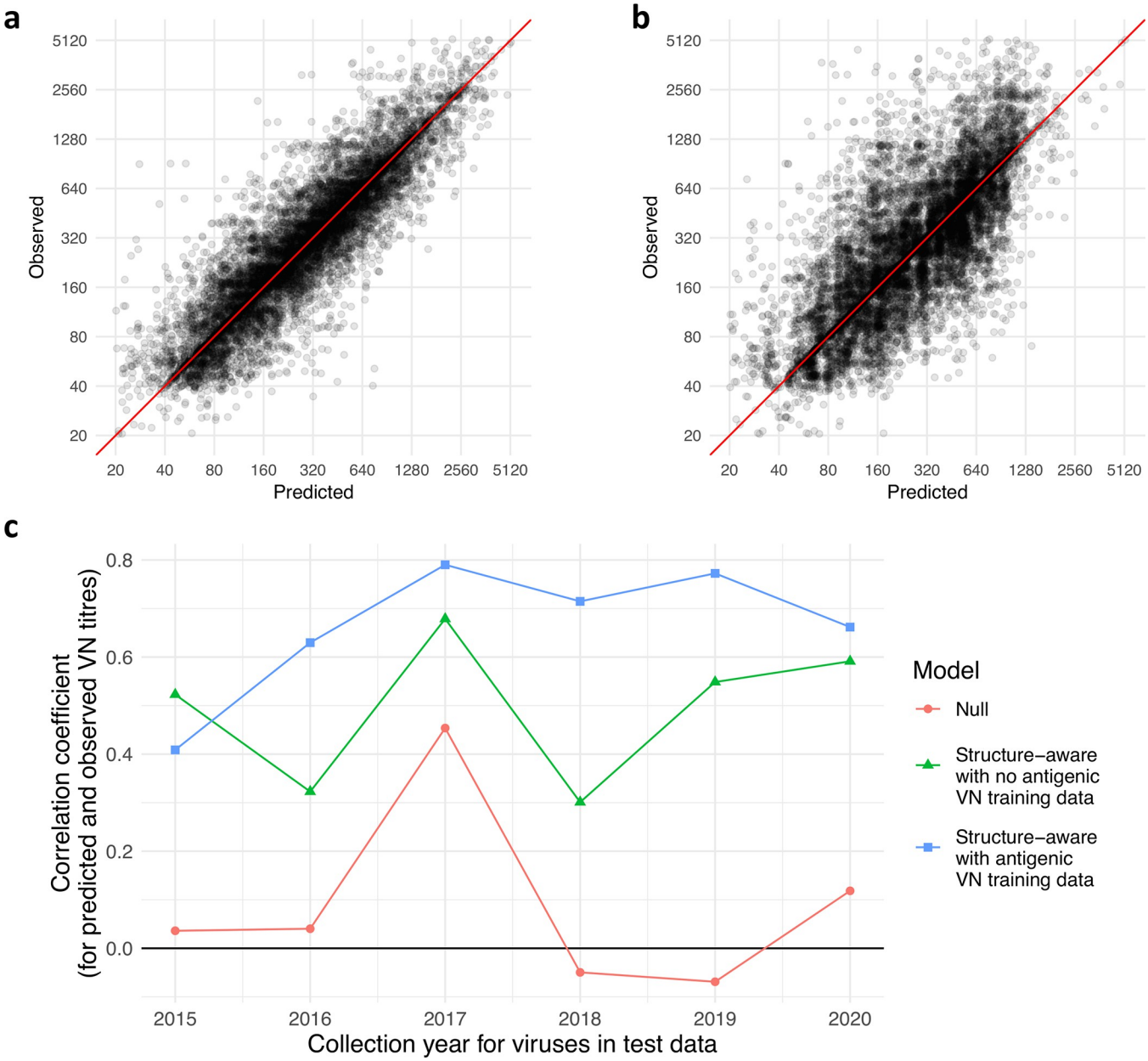

Fig 5. Sequence-based predictions of antigenic phenotype. Measured HI titres are plotted against predictions made under cross-validation procedures where test sets consist of randomly selected (**a**) 10% of test-reference virus pairs and (**b**) 10% of test viruses. Predictions are made using a structurally-aware model (Eq 6) in which structural features associated with the location of each residue in the HA protein influence the combination of genetic terms that contribute to predictions. Observed titres are the fitted titre for each reference strain and test virus, accounting for day-to-day variability in measured titres. (**c**) Pearson's correlation coefficient between observed titres and predictions made for test sets consisting of VN measurements between viruses collected in a specified year and antisera present in training datasets of viruses collected in previous years. The null model estimates titres based on homologous titre values for antisera in the training dataset only. The structural model with no antigenic information from the training VN data estimates titres using homologous titre values for an antiserum in the training dataset and structure-informed probabilities (π_λ in Eq 7) and position-specific antigenic terms ($\tilde{m}_\lambda$ in Eq 6) from a structurally-aware model applied to HI data collected from 1990–2014 (summarised in Fig 4). The structurally-aware model with antigenic information is additionally able to train estimates of antigenic effects of substitutions present in VN training datasets.

https://doi.org/10.1371/journal.pcbi.1010885.g005

genetic signals which can prove challenging for accurate mapping of genotype to antigenic phenotype. We describe a Bayesian approach whereby the causative amino acid changes are identified using BSSVS, performed alongside model fitting. Incorporating phylogenetic

structure into the analysis favours amino acid positions where there are substitutions correlating with antigenic change at multiple points in the evolutionary history of the virus [12]. Moreover, we show how information on the structural context of amino acid positions can guide the selection of antigenically relevant genetic variables for the HA of influenza A(H3N2) viruses, increasing the proportion of variables either included or excluded with high confidence (from 59.8% to 72.4%) and increasing the tendency for the model to attribute antigenic variation to substitutions at HA positions within or nearby recognised antigenic sites (reducing the mean distance across samples from 6.86 Å to 2.55 Å). We identify substitutions responsible for antigenic changes noting remarkable repeatability in the role of key positions and describing instances of large antigenic changes that proved evolutionarily successful and otherwise. Accurately quantifying the impact of substitutions on phenotype and monitoring such substitutions as they arise in different contexts should allow for a fuller understanding of how different aspects of phenotype contribute to viral fitness. Such steps are essential in the quest to understand the forces that govern the predictability of evolution.

A benefit of a BSSVS approach with indicator variables is that this explicitly accounts for uncertainty in the identification of causative amino acid substitutions and reflects this uncertainty in variable selection. A further advantage of a Bayesian approach in this context is the opportunity to include prior information on the antigenicity of HA, here in the form of structural data; something we show can help to further resolve ambiguities in model selection. We describe an approach to integrate data on proximity to the RBS and a pre-computed structure-based epitope score into priors for the probability that substitutions at a position impact antigenicity, an approach that could be extended to other proteins or structural features. In this analysis of the influenza HA, a prior expectation that antigenically important substitutions would tend to occur at HA positions both increased proximity to the RBS and higher epitope scores was implemented through choice of priors. However the model was free to minimise the influence of either parameter and this approach is not limited to situations in which prior knowledge of the directionality of a relationship exists. If a misleading or unhelpful structural feature is selected prior to the analysis, the posterior distribution for the associated power term is expected to tend towards zero minimising the influence of the feature (**S2 and S3 Figs**). An extension of this work would be to account for the shielding of epitopes by covalently attached glycans and the changes in the accessibility of epitopes that occurs as a result of changes in glycosylation over time [25,26]. Our analysis was focused on the HA protein as the HI assay measures antigenic variation in HA. However, the analysis could be extended to also include neuraminidase to test for an antigenic effect of substitutions in that glycoprotein which may influence VN assays. There is also scope to exploit the transfer of information from data-rich to data-starved virus subtypes, similar to that performed in the prediction scheme for VN titres for viruses evolved in the years following the period covered by the principal dataset analyses in this study (**Fig 5c**). For example, the relationship between structural features and the role in antigenic evolution could be trained using a data-rich subtype such as A(H3N2); this could inform the prior distributions for the inclusion probabilities of structurally-aligned positions in any newly emerged subtype that would have fewer available data.

Having developed the genetic model of antigenicity to fit known data, its predictive power was evaluated by cross-validation using multiple schemes. Under the first, titres were predicted for training datasets comprising 10% of virus and antiserum combinations. As much as 87% of predicted titres were within 1 $\log_2$ HI unit of the observed titre and 100% within 2 $\log_2$ HI units and **Fig 5a** shows accuracy across all values of observed titre. This would mean, for example, that if we knew some of the characteristics of a test virus we could accurately estimate its HI titre against a new reference antiserum, a vaccine virus antiserum or a candidate vaccine virus antiserum within a two-fold dilution in most cases. The accuracy of the prediction was

improved by the inclusion of variables representing amino acid substitutions (**Eqs 4–6**) over the basal phylogenetic model (**Eq 3**) which, lacking terms representing specific amino acid differences, cannot recognise when clades separated on the tree are antigenically similar because they share substitutions occurring convergently in different branches. Predictions were not markedly enhanced by linking the magnitude of antigenic effects caused by different amino acid substitutions occurring at the same HA position (**Eq 5**), or by the inclusion of structural information to guide the identification of the substitutions (**Eq 6**).

In the second, more challenging, prediction scheme we removed all antigenic data for a virus and predicted its titres against reference antisera. Here, predictions were somewhat less accurate though up to 71% of predicted titres were within 1 $\log_2$ HI unit of the observed titre and 94% within 2 $\log_2$ HI units. In this prediction scheme (scheme 2) shown in **Fig 5b**, it is clear that the model rarely predicts high HI titres (e.g. >1280) with no knowledge of whether the 'missing' test virus has an inherent ability to be more sensitive to specific or non-specific inhibition by antisera. This result would, in the absence of any established antigenic data, predict whether an unknown virus was not well recognised by an antiserum (recognition of the test virus by antisera raised against reference viruses at titres 4-fold lower than the titre with the homologous virus) or was poorly recognised by such antisera (recognition at titres >4-fold lower than the homologous titres). Again, the greatest enhancement was, as seen in scheme 1, due to the inclusion of specific amino acid substitutions over and above the basal phylogenetic model. However, in scheme 2, the accuracy of the predictions improved, first by inclusion of specific amino acid substitutions (**Eq 4**), then by linkage of antigenic weights of alternative substitutions at a position (**Eq 5**), and finally by including information on the proximity of residues to the RBS and structure-based epitope scores (**Eq 6**).

In summary, the ability to predict antigenic cross-reactivity of emerging influenza viruses, as measured by HI, while maximising identification of the causative amino acid substitutions provides important information with which to evaluate the epidemic potential of influenza virus variants. The development of accurate, quantitative genotype-to-phenotype maps are a required step towards the development of accurate sequence-based prediction of viral fitness and genotype emergence and success, a hugely exciting area for further research. Using models parameterised using data collected in previous years can help to refine such techniques and we describe how structural data can be incorporated into model fitting. Incorporating other sources of prior information is an exciting area for further model development, for example alternative protein structural data could be tested as explaining variation in alternative assays used to characterise antigenic similarity of viruses. The benefits of being able to integrate such data types into modelling of evolution could be particularly powerful when performing analyses of emerging influenza viruses for which historic data are unavailable. The approach we describe, allowing detailed and accurate mapping of genotype to antigenic phenotype, should progress efforts to understand the genetic determinants of virus fitness and evolutionary trajectories of influenza viruses, importantly when surveillance is increasingly based on a 'sequence-first' approach. Moreover, this approach could also be adapted to proteins of other viruses such as the capsid proteins of FMDV and the spike protein of coronaviruses.

Materials and methods

Influenza data

Influenza viruses were originally isolated from clinical specimens either by WHO National Influenza Centres or by the London-based WHO Collaborating Centre (CC). The antigenic dataset for A(H3N2) included 1737 viruses for which HI and HA1-encoding gene sequence data were generated at the CC. The VN dataset consisted of 8,734 titres carried out between

5,325 combinations of virus and antiserum. All HI and VN data used were from assays performed at the CC and were obtained using post-infection ferret antisera. The data associated with this study are available online [27]. HA1 nucleotide sequences and collection dates were analysed to generate temporal phylogenies using BEAST v1.8.2 [28]. Phylogenies were estimated using a variety of nucleotide substitution, demographic, and molecular clock models. A general time reversible model of nucleotide substitution with proportion of invariant sites and a gamma distribution with four categories describing among-site variation (GTR + I + Γ_4) was determined to be the most suitable model by comparison of Bayes factors. The trunk lineage was defined from the root through the descendant node leading to the greater number of sampled viruses.

Structural analysis

For each residue in the structure, the distance from the RBS was calculated as the minimum distance in 3-D space between the alpha carbon of that residue and the nearest alpha carbon of a residue in the RBS (positions 98, 135, 136, 153, 183, 190, 194, 195, 225, 226, and 228) [29]. The distance of each HA residue to the nearest antigenic site constituent was calculated as the minimum distance in 3-D space from the residue's alpha carbon to the nearest alpha carbon of a residue in the antigenic site A (positions 124, 126, 128, 131, 133, 135, 137, 142, 143, 144, 145, and 146), site B (155, 156, 157, 158, 159, 160, 163, 164, 186, 188, 189, 190, 192, 193, 196, 197, 198 and 199), site C (48, 53, 54, 275, 276, 278), site D (121, 122, 171, 172, 173, 174, 207, 217, 226, 227, 242, 244 and 248) and site E (62, 63, 78, 81, 82, 83, 91, 92, 260 and 262) [22,30,31]. To determine structure-based epitope scores for each residue in the HA structure from tertiary structure, the program BEpro [23] was used to analyse structures in Protein data bank (PDB) format. These scores reflect side chain orientation and solvent accessibility calculated using half sphere exposure values at multiple distances and amino acid propensity scores. For each residue, both half sphere exposure measures and propensity scores depend on all atoms within 8-16Å of the target residue, with increased weighting towards nearer atoms. Due to this, scores for any given residue are relatively insensitive to the effects of single amino acid substitutions. The BEpro server was accessed at http://pepito.proteomics.ics.uci.edu/info.html. Structural features were calculated from the solved HA structures of three human A(H3N2) viruses isolated during the period covered by the HI dataset, A/Finland/486/2004 (PDB: 2YP2 [32]), A/Hong Kong/4443/2005 (PDB: 2YP7 [32]), and A/Brisbane/10/2007 (PDB: 6AOU [33]) and from an earlier virus isolated in 1968 (A/Aichi/2/68 PDB: 3HMG [34]).

Model implementation

The *Modelling Approach* section above describes the modelling process by which variation in HI titres was attributed to genetic differences between viruses while accounting for phylogenetic relationships and non-antigenic sources of variation in titres using a hierarchical model structure. The terms used in the models are described in **Table 1**. This section describes choices of prior and other implementation details. Models were fitted in JAGS v4.3.0 [35] using the R package runjags v2.0.4–6 [36]. Post-infection ferret antisera were raised against a range of reference viruses (r) and HI titres measured for an individual antiserum and several viruses (v) (including the homologous titre to the corresponding reference virus r and a range of other viruses). The measured titre for antiserum against reference virus r and virus v on a given date d, $Y_{r,v,d}$ was modelled as an underlying titre for the combination of reference virus and test virus pair, $H_{r,v}$, and an effect for date along with a variance term (**Eq 1**). To reflect a lack of prior information, the effect for date was implemented with a diffuse prior for the effect of date defined as a normal distribution where $D_d \sim \text{Normal}(0, \sigma_D^2)$ with the variance parameter σ_D^2

was drawn from an inverse gamma distribution such that $\sigma_D^2 \sim \text{Inverse Gamma}(0.001, 0.001)$. The variance term accounting for residual variation in measured titres was drawn from the prior distribution $\sigma_Y^2 \sim \text{Inverse Gamma}(0.001, 0.001)$. The underlying titre for each test virus and reference strain was modelled according to a general structure described by Eq 2. Priors for the estimated impact on titres associated with the use of antiserum raised using each reference virus were defined as $I_r \sim \text{Normal}(\mu_I, \sigma_I^2)$. Here, the mean was given the prior $\mu_I \sim \text{Normal}(Y_{max}, \sigma_{\mu_I}^2)$ where Y_{max} was the maximum observed titre and $\sigma_{\mu_I}^2 = 1.5$ to maintain a value for the intercept proximal to the range of possible recorded titres without explicitly enforcing a particular value such as the maximum observed titre within the dataset and $\sigma_I^2 \sim \text{Inverse Gamma}(0.001, 0.001)$. The prior for μ_I was chosen to allow a fitted model with a high intercept from which lower titres were fitted by the subtraction of positive antigenic terms. Effects associated with the use of each test virus were associated with the prior $A_v \sim \text{Normal}(0, \sigma_A^2)$ where $\sigma_A^2 \sim \text{Inverse Gamma}(0.001, 0.001)$. The variance term accounting for residual variation in underlying titres was drawn from the prior distribution $\sigma_H^2 \sim \text{Inverse Gamma}(0.001, 0.001)$.

The antigenic component of the model described in Eq 3, $\Delta_{r,v}$, estimates the drop in HI titres caused by the antigenic dissimilarity of reference strain r and test virus v. $\Delta_{r,v}$ was modelled in several ways (**Eqs 3–6**) though was restricted to be positive throughout and was always subtracted from the other terms contributing to variation in titres (Eq 3). Using the combination of branches of the phylogenetic tree traversed when a path between them is drawn through the tree (Eq 3), each branch, ψ, of the phylogenetic tree, Ψ, was associated with a binary mask term, γ_ψ, that determined whether the branch ψ was included in the model or not and an antigenic weight parameter, m_ψ, representing the estimated drop in HI titres due to antigenic change mapping to the branch. These antigenic weights were required to be non-negative and the prior was defined as $m_\psi \sim \text{Gamma}(2, 1)$. This choice of prior distribution discouraged the inclusion in the model of a high number of branches associated with very small antigenic weights, on the basis that effects very close to zero cannot be identified in HI assay data, thereby encouraging a more parsimonious model. Binary mask variables associated with each branch were drawn from a Bernoulli trial defined as $\gamma_\psi \sim \text{Bernoulli}(\pi)$ where π was given the prior $\pi \sim \text{Beta}(2,8)$ to favour a sparse model reflecting the expectation that a relatively low proportion of branches were expected to represent genetic differences influencing titres. The posterior mean value of π determined the proportion of branches used to account for phylogenetic structure when testing variables representing specific amino acid differences (**Eqs 4, 5 and 6**).

Next, terms were introduced to explicitly attribute antigenic differences between viruses expressed in HI assays to specific amino acid differences (Eq 4). These terms representing specific amino acid difference were tested in the presence of a subset of phylogenetic terms selected using Eq 3, $\hat{\Psi}$. Each amino acid position was associated with a binary mask term, ζ_λ, for which the prior was specified as $\zeta_\lambda \sim \text{Bernoulli}(\pi')$ where given the prior $\pi' \sim \text{Beta}(2,8)$ mirroring that used for the binary mask associated with phylogenetic terms above. When implementing Eq 4, priors for the effect sizes associated with substitutions were specified as $m'_\kappa \sim \text{Gamma}(1, 1)$, which allows for antigenic effects close to zero—necessary as this allows for substitutions that do not affect antigenic cross-reactivity to occur at amino acid positions included in the model due to the presence of other antigenically important substitutions at the position. A similar model where substitutions occurring at the same amino acid position are linked in their effect size is described by Eq 5, which differs from Eq 4 by having the antigenic effect of a substitution partitioned into position-specific and substitution-specific components, $\tilde{m}_\lambda$ and $\tilde{m}\prime_\kappa$ respectively. Priors for the position-specific antigenic parameter were specified as

$\tilde{m}_{\lambda} \sim \text{Gamma}(2, 1)$, discouraging near-zero antigenic effects at the position level. Eq 5 was implemented with the prior for specific amino acid substitutions specified as $\tilde{m}\prime_{\kappa} \sim \text{Gamma}(\alpha, \alpha)$ where α~Gamma(2, 1). This allows for a range of small and large antigenic impacts, but the mean of this prior is equal to 1 regardless of the value of α. Therefore, for rarer substitutions (ones informed by very few titres) where the data make reliable estimation of an effect size unlikely, $\tilde{m}\prime_{\kappa}$ tends towards 1, so the antigenic impact of the substitution will be largely determined by the value estimated for the position-specific antigenic effect $\tilde{m}_{\lambda}$.

Structural data associated with each amino acid position were used to influence the probability that substitutions at that position contribute to antigenic evolution as apparent in HI titres (**Eqs 7 and 8**). The binary mask term for each position, $\tilde{\zeta}_{\lambda}$, now depends on a position-specific structure-based probability of antigenic importance π_{λ}. Two features derived from structural analysis of published H3 HA structures, as described above, were defined for each HA position: the distance from the RBS, F_1, and a structure-based epitope score, F_2 (see Structural analysis section). Each of these were re-scaled between zero and one prior to modelling so that higher values reflected high proximity to the RBS and high epitope scores respectively. Whether or not substitutions at an HA position contributed to variation in HI titres depended on the outcome of a Bernoulli trial where for each position, λ, a structure-informed probability π_{λ} was determined as the product of $F_{1,\lambda}$ and $F_{2,\lambda}$. Higher values of each structural feature were expected to increase the probability that a position was antigenically important and so the parameters $\rho 1$ and $\rho 2$ were restricted to be positive and their priors were defined as ρ_1~Gamma(1, 1) and ρ_2~Gamma(1, 1).

To give a sense of the complexity of the models described and computational requirements, when the structurally-naïve with substitution effect sizes linked to position (Eq 5) is applied to the full dataset of 38,758 titres (observed stochastic nodes) there are 14,533 unobserved stochastic nodes if forward and reverse substitutions are assumed to have symmetric antigenic effects and 14,821 if they are not. This gave total graph sizes of 3,040,790 and 3,043,157 respectively. Implemented with 5,000 step MCMC chains after 1,000 adaptive iterations, two chains took approximately 40 and 44.5 hours to run on a desktop dual-processor Ubuntu 20.04 Linux workstation (with 18-core 2.3GHz processors and 256 GB RAM), respectively.

Cross-validation using Bayesian model averaging

Two cross-validation schemes were performed with the full dataset repeatedly divided into training and test datasets at random, criteria for each are described in the Results section. Under each scheme, measurements for antisera not present in the training data were excluded from the test dataset. Models described by **Eqs 3–6** were each fitted to the training data and used to predict HI titres for virus and antiserum combinations present in the test data. Predicted titres were compared with observed titres, with both MAE and RMSD calculated. Each error influences MAE in direct proportion to the absolute value of the error whereas RMSD places more emphasis on penalisation of higher errors. For the forward prediction scheme carried out with VN data, test datasets consisted of the following number of viruses in each year: 2015 (76); 2016 (250); 2017 (226); 2018 (248); 2019 (222); and 2020 (62).

Supporting information

S1 Fig. Correlation between structural features of A/Aichi/2/68 and A/Brisbane/10/2007 HA. (**a**) The correlation in the distance of the alpha carbon of each HA residue to the closest alpha carbon of a residue belonging to a described antigenic site. (**b**) The correlation in the distance of the alpha carbon of each HA residue to the closest alpha carbon of a residue belonging to the receptor-binding site. (**c**) The correlation in structure-based epitope scores estimated for

each HA residue using the software BEpro.
(TIF)

S2 Fig. Posterior distributions for structural model provided with proximity to the HA1-HA2 cleavage site and a randomised score and a structure-based epitope score. (**a**) The distance of each HA residue to the HA1-HA2 boundary where post-translational cleavage occurs. A vertical dashed line at position 329 indicates the boundary between HA1 and HA2. To the right, a surface representation of the HA is shown. (**b**) In each plot, posterior distributions for power terms that link proximity to the cleavage site (left) and structure-based epitope scores (right) for each HA position to a structure-informed probability for the position, π_λ, according to **Eq 8.** Individual values sampled from the posterior distribution are shown below the x-axis. Prior distributions for these parameters, defined as Gamma(1, 1), are shown as dashed black lines.
(TIF)

S3 Fig. Posterior distributions for structural model provided with a randomised score and proximity to receptor-binding site. (**a**) Values randomly drawn from a beta(1,1) distribution and assigned to each HA residue. A vertical dashed line at position 329 indicates the boundary between HA1 and HA2. To the right, a surface representation of the HA is shown. (**b**) In each plot, posterior distributions for power terms that link randomly drawn values (left) and proximity to the RBS (right) for each HA position to a structure-informed probability for the position, π_λ, according to **Eq 8.** Individual values sampled from the posterior distribution are shown below the x-axis. Prior distributions for these parameters, defined as Gamma(1, 1), are shown as dashed black lines.
(TIF)

S1 Table. Model confidence in variable selection and distance of included or excluded HA positions to known antigenic sites.
(DOCX)

S1 Data. Data underlying Fig 2 which summarises results of structurally-naïve model.
(CSV)

S2 Data. Data underlying Fig 3 which summarises structural features of influenza A (H3N2) HA.
(CSV)

S3 Data. Data underlying Fig 4 which summarises results of structurally-aware model.
(CSV)

Acknowledgments

We acknowledge the network of WHO National Influenza Centres and WHO Collaborating Centres that comprise the WHO Global Influenza Surveillance and Response System, who provided viruses described in datasets analysed in this study.

Author Contributions

Conceptualization: William T. Harvey, Rodney S. Daniels, Alan J. Hay, John W. McCauley, Richard Reeve.

Data curation: William T. Harvey, Rodney S. Daniels, Lynne Whittaker, Victoria Gregory, Richard Reeve.

Formal analysis: William T. Harvey, Richard Reeve.

Funding acquisition: Rodney S. Daniels, John W. McCauley, Richard Reeve.

Investigation: Lynne Whittaker, Victoria Gregory.

Methodology: William T. Harvey, Vinny Davies, Dirk Husmeier, Richard Reeve.

Project administration: John W. McCauley, Richard Reeve.

Resources: Rodney S. Daniels, John W. McCauley.

Software: William T. Harvey, Vinny Davies, Dirk Husmeier, Richard Reeve.

Supervision: Rodney S. Daniels, Alan J. Hay, Dirk Husmeier, John W. McCauley, Richard Reeve.

Visualization: William T. Harvey.

Writing – original draft: William T. Harvey.

Writing – review & editing: William T. Harvey, Rodney S. Daniels, John W. McCauley, Richard Reeve.

References

1. WHO. Influenza (seasonal) fact sheet. In: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). 2018.
2. Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet. 2018; 391: 1285–1300. https://doi.org/10.1016/S0140-6736(17)33293-2 PMID: 29248255
3. Monto AS. Reflections on The Global Influenza Surveillance and Response System (GISRS) at 65 Years: An Expanding Framework for Influenza Detection, Prevention and Control. Influenza Other Respi Viruses. 2018; 12: 10–12. https://doi.org/10.1111/irv.12511 PMID: 29460424
4. Hay AJ, McCauley JW. The Global Influenza Surveillance and Response System (GISRS)—A Future Perspective. Influenza Other Respi Viruses. 2018; 0–2. https://doi.org/10.1111/irv.12565
5. Morris DH, Gostic KM, Pompei S, Bedford T, Łuksza M, Neher RA, et al. Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. Trends in Microbiology. 2018. pp. 102–118. https://doi.org/10.1016/j.tim.2017.09.004 PMID: 29097090
6. Neher RA, Russell CA, Shraiman BI. Predicting evolution from the shape of genealogical trees. Elife. 2014; 3: e03568. https://doi.org/10.7554/eLife.03568 PMID: 25385532
7. Łuksza M, Lässig M. A predictive fitness model for influenza. Nature. 2014; 507: 57–61. https://doi.org/10.1038/nature13087 PMID: 24572367
8. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci. 2016; 113: E1701–E1709. https://doi.org/10.1073/pnas.1525578113 PMID: 26951657
9. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus ADME, et al. Mapping the antigenic and genetic evolution of influenza virus. Science. 2004; 305: 371–376. https://doi.org/10.1126/science.1097211 PMID: 15218094
10. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, et al. Integrating influenza antigenic dynamics with molecular evolution. Elife. 2014; 3: e01914. https://doi.org/10.7554/eLife.01914 PMID: 24497547
11. Sun H, Yang J, Zhang T, Long L, Jia K, Yang G, et al. Using sequence data to infer the antigenicity of influenza virus. MBio. 2013; 4: e00230–13. https://doi.org/10.1128/mBio.00230-13 PMID: 23820391
12. Harvey WT, Benton DJ, Gregory V, Hall JPJ, Daniels RS, Bedford T, et al. Identification of Low- and High-Impact Hemagglutinin Amino Acid Substitutions That Drive Antigenic Drift of Influenza A(H1N1) Viruses. Hensley SE, editor. Pathog PLOS. 2016; 12: e1005526. https://doi.org/10.1371/journal.ppat.1005526 PMID: 27057693
13. Reeve R, Blignaut B, Esterhuysen JJ, Opperman P, Matthews L, Fry EE, et al. Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus.

Tanaka MM, editor. PLoS Comput Biol. 2010; 6: e1001027. https://doi.org/10.1371/journal.pcbi.1001027 PMID: 21151576

14. Peacock TP, Harvey WT, Sadeyen J, Reeve R, Iqbal M. The molecular basis of antigenic variation among A(H9N2) avian influenza viruses. Emerg Microbes Infect. Springer US; 2018; 7: 176. https://doi.org/10.1038/s41426-018-0178-y PMID: 30401826
15. Davies V, Reeve R, Harvey WT, Maree FF, Husmeier D. A sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution. Comput Stat. Springer Berlin Heidelberg; 2017; 32: 803–843. https://doi.org/10.1007/s00180-017-0730-6
16. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996; 58: 267–288.
17. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005; 67: 301–320.
18. Davies V, Harvey WT, Reeve R, Husmeier D. Improving the identification of antigenic sites in the H1N1 influenza virus through accounting for the experimental structure in a sparse hierarchical Bayesian model. J R Stat Soc Ser C (Applied Stat. 2019; https://doi.org/10.1111/rssc.12338 PMID: 31598013
19. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging. Stat Sci. 1999; 14: 382–401.
20. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006; 4: e88. https://doi.org/10.1371/journal.pbio.0040088 PMID: 16683862
21. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018; 4: 1–5. https://doi.org/10.1093/ve/vey016 PMID: 29942656
22. Wiley DC, Skehel JJ. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu Rev Biochem. 1987; 56: 365–394. https://doi.org/10.1146/annurev.bi.56.070187.002053 PMID: 3304138
23. Sweredoski MJ, Baldi P. PEPITO: Improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure. Bioinformatics. 2008; 24: 1459–1460. https://doi.org/10.1093/bioinformatics/btn199 PMID: 18443018
24. Koel BF, Burke DF, Bestebroer TM, van der Vliet S, Zondag GCM, Vervaet G, et al. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science. 2013; 342: 976–979. https://doi.org/10.1126/science.1244730 PMID: 24264991
25. Blackburne BP, Hay AJ, Goldstein RA. Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog. 2008; 4: e10000058. https://doi.org/10.1371/journal.ppat.1000058 PMID: 18451985
26. Altman MO, Angel M, Košík I, Trovão NS, Zost SJ, Gibbs JS, et al. Human influenza a virus hemagglutinin glycan evolution follows a temporal pattern to a glycan limit. MBio. 2019; 10: e00204–19. https://doi.org/10.1128/mBio.00204-19 PMID: 30940704
27. Whittaker L, Gregory V, Harvey WT, Daniels RS, Reeve R, Halai C, Douglas A, Gonsalves R, Skehel JJ, Hay AJ and McCauley JW (2023) Human seasonal Influenza A(H3N2) haemagglutination inhibition data 1990–2021 from the WHO Collaborating Centre for Reference and Research on Influenza, London, UK. https://doi.org/10.5525/gla.researchdata.1405
28. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012; 29: 1969–1973. https://doi.org/10.1093/molbev/mss075 PMID: 22367748
29. Skehel JJ, Wiley DC. Receptor binding and membrane fusion in virus entry: The influenza hemagglutinin. Annu Rev Biochem. 2000; 69: 531–569. https://doi.org/10.1146/annurev.biochem.69.1.531 PMID: 10966468
30. Wiley DC, Wilson IA, Skehel JJ. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature. 1981; 289: 373–378. https://doi.org/10.1038/289373a0 PMID: 6162101
31. Wilson I, Cox N. Structural basis of immune recognition of influenza virus hemagglutinin. Annu Rev Immunol. 1990; 8: 737–771. https://doi.org/10.1146/annurev.iy.08.040190.003513 PMID: 2188678
32. Lin YP, Xiong X, Wharton SA, Martin SR, Coombs PJ, Vachieri SG, et al. Evolution of the receptor binding properties of the influenza A (H3N2) hemagglutinin. Proc Natl Acad Sci U S A. 2012; 109: 21474–21479. https://doi.org/10.1073/pnas.1218841110 PMID: 23236176
33. Wu NC, Zost SJ, Thompson AJ, Oyen D, Nycholat CM, McBride R, et al. A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. PLoS Pathog. 2017; 13: 1–17. https://doi.org/10.1371/journal.ppat.1006682 PMID: 29059230
34. Weis WI, Brünger AT, Skehel JJ, Wiley DC. Refinement of the influenza virus hemagglutinin by simulated annealing. J Mol Biol. 1990; 212: 737–761. https://doi.org/10.1016/0022-2836(90)90234-D PMID: 2329580

35. Plummer M. Just Another Gibbs Sampler v3.3.0 (JAGS): A program for analysis of Bayesian graphical models using Gibbs sampling. http://mcmc-jags.sourceforge.net. 2012.

36. Denwood MJ. runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions For MCMC Models in JAGS. J Stat Softw. 2016;71. https://doi.org/10.18637/jss.v071.i09