AMERICAN SOCIETY FOR MICROBIOLOGY mSphere®

RESEARCH ARTICLE

Machine Learning Prediction and Experimental Validation of Antigenic Drift in H3 Influenza A Viruses in Swine

Michael A. Zeller,[a,b] Phillip C. Gauger,[a] Zebulun W. Arendsee,[c] Carine K. Souza,[c] Amy L. Vincent,[c] Tavis K. Anderson[c]

[a]Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University, Ames, Iowa, USA
[b]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, USA
[c]Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, Iowa, USA

ABSTRACT The antigenic diversity of influenza A viruses (IAV) circulating in swine challenges the development of effective vaccines, increasing zoonotic threat and pandemic potential. High-throughput sequencing technologies can quantify IAV genetic diversity, but there are no accurate approaches to adequately describe antigenic phenotypes. This study evaluated an ensemble of nonlinear regression models to estimate virus phenotype from genotype. Regression models were trained with a phenotypic data set of pairwise hemagglutination inhibition (HI) assays, using genetic sequence identity and pairwise amino acid mutations as predictor features. The model identified amino acid identity, ranked the relative importance of mutations in the hemagglutinin (HA) protein, and demonstrated good prediction accuracy. Four previously untested IAV strains were selected to experimentally validate model predictions by HI assays. Errors between predicted and measured distances of uncharacterized strains were 0.35, 0.61, 1.69, and 0.13 antigenic units. These empirically trained regression models can be used to estimate antigenic distances between different strains of IAV in swine by using sequence data. By ranking the importance of mutations in the HA, we provide criteria for identifying antigenically advanced IAV strains that may not be controlled by existing vaccines and can inform strain updates to vaccines to better control this pathogen.

IMPORTANCE Influenza A viruses (IAV) in swine constitute a major economic burden to an important global agricultural sector, impact food security, and are a public health threat. Despite significant improvement in surveillance for IAV in swine over the past 10 years, sequence data have not been integrated into a systematic vaccine strain selection process for predicting antigenic phenotype and identifying determinants of antigenic drift. To overcome this, we developed nonlinear regression models that predict antigenic phenotype from genetic sequence data by training the model on hemagglutination inhibition assay results. We used these models to predict antigenic phenotype for previously uncharacterized IAV, ranked the importance of genetic features for antigenic phenotype, and experimentally validated our predictions. Our model predicted virus antigenic characteristics from genetic sequence data and provides a rapid and accurate method linking genetic sequence data to antigenic characteristics. This approach also provides support for public health by identifying viruses that are antigenically advanced from strains used as pandemic preparedness candidate vaccine viruses.

KEYWORDS antigenic drift, influenza A, machine learning, molecular epidemiology, swine, viral evolution

Citation Zeller MA, Gauger PC, Arendsee ZW, Souza CK, Vincent AL, Anderson TK. 2021. Machine learning prediction and experimental validation of antigenic drift in H3 influenza A viruses in swine. mSphere 6:e00920-20. https://doi.org/10.1128/mSphere.00920-20.

Editor Seema Lakdawala, University of Pittsburgh School of Medicine

Address correspondence to Tavis K. Anderson, tavis.anderson@usda.gov.

Machine learning prediction of antigenic drift in H3 swine influenza A viruses. @ISUVetMed @bcbiastate @Tavis_Anderson

Received 9 September 2020
Accepted 23 February 2021
Published 17 March 2021

Influenza A virus (IAV) is a primary respiratory pathogen in commercial swine in the United States (1). Preventing infection and transmission of the virus has proven difficult due to rapid mutation that allows the virus to evade host immune defenses and impacts the efficacy of vaccination programs by antigenic drift (2). The best approach for effective IAV control has been the development of vaccines that reflect the antigenic diversity of circulating swine IAV strains (3). This is dependent on robust sampling and sequencing of contemporary strains, which is currently achieved primarily through passive surveillance, whereby clinically sick pigs are sampled and the hemagglutinin (HA) gene is sequenced and compared to vaccine antigens based on either genetic clade or sequence identity. Vaccines that include a well-matched HA can induce the production of antibodies that may provide sterilizing immunity, help reduce clinical signs, or reduce transmission (4, 5). Conversely, mismatched vaccine antigens can result in vaccine failure or potentially cause enhanced disease, emphasizing the importance of careful vaccine strain selection (6).

In the United States, swine IAV is monitored by the U.S. Department of Agriculture (USDA) in collaboration with regional veterinary diagnostic laboratories in the National Animal Health Laboratory Network (7). These data are synthesized primarily using phylogenetic analysis (7, 8), but there is no coordinated effort to characterize the phenotypic differences between circulating viruses (9). This contrasts with the approach for human IAV, whereby vaccine antigens are selected through comprehensive genetic and antigenic characterization of seasonally circulating IAV strains (10). Thus, the majority of vaccine antigens in use for IAV in swine are selected based solely on the genetic clade or amino acid identity. This effort is fraught with risk, as there are at least 16 distinct HA genetic clades of IAV in swine derived from multiple human-to-swine interspecies transmission events and subsequent evolution in the swine host (8, 11). Further, there is evidence for regional patterns in HA clade persistence (8, 12) and as few as six amino acid mutations within the HA may affect the antigenic phenotype of a virus (13, 14). Consequently, there is a critical need to not only sequence and genetically characterize swine IAV but also determine what of the genetic diversity is meaningful for antigenic drift.

The antigenic properties of IAV are a manifestation of the structural interaction between IAV and host antibodies (15–18). Structural changes in the HA may alter the interaction with antibodies targeting the virus, and these changes are generally correlated with the number of accumulated amino acid mutations in the HA protein (19). Empirical data have also shown that certain amino acid mutations have a disproportionate effect on antigenic change based on the location of the amino acid in the protein structure (13, 15). Though there are relatively few antigenically characterized swine IAV HA genes (9, 13), these empirical data may be used to establish antigenic distances between multiple IAVs in swine and to gain insight into the contribution of site-specific amino acid mutations. These data can subsequently be used to predict antigenic drift and assign a ranking of importance to specific amino acid mutations that nuance the biological relevance of genetic diversity collected during surveillance programs.

In this study, machine learning methods were used to model the antigenic properties of IAV in swine and predict the antigenic distance between different strains using HA sequences. Modeling methods, such as the ones we present, are able to overcome the prohibitive costs and logistical challenges associated with large-scale phenotypic characterization. These data can be used in combination with in-field surveillance platforms (20) as an approach for the early detection of antigenic variants and novel viruses. Additionally, these algorithms can be disseminated to swine practitioners in analytical pipelines (11, 20, 21) to facilitate the rational design of vaccines that include antigens that will likely protect against the circulating IAV strains. Understanding how genetic diversity, and which amino acids within the HA gene are the most important, can allow for the simulation of the antigenic evolution of swine IAV and make predictions about the persistence and circulation of future IAV strains.

TABLE 1 Performance indicators for the random forest, AdaBoost decision tree, multilayer perceptron, and ensemble regression models with tuned hyperparameters[a]

	Value by indicated model			
Performance indicator	Random forest	AdaBoost decision tree	Multilayer perceptron	Ensemble
Pearson correlation	0.78	0.77	0.78	0.80
RMSE	1.60	1.28	1.32	1.21
10-fold CV (RMSE)	1.56 (±0.29)	1.59 (±0.33)	1.76 (±0.39)	1.58 (±0.27)

[a] Pearson correlation and root mean square error (RMSE) were determined using an 80%/20% split between training and test antigen data. A 10-fold cross validation based on the RMSE was applied. CV, cross validation.

RESULTS

Machine learning model performance. Comparison of the empirical antigenic distances with the values predicted by random forest, AdaBoost decision tree, multilayer perceptron regression, and the ensemble of all three models indicated that the Pearson correlation for all regression models was within a range of 77% to 80% (Table 1). The root mean square error (RMSE) was between 1.21 and 1.60 antigenic units (AU) of error depending on the model. Tenfold cross validation of the random forest, AdaBoost decision tree, multilayer perceptron, and the ensemble of the regression models had RMSEs of 1.56 ± 0.29, 1.59 ± 0.33, 1.76 ± 0.39, and 1.58 ± 0.27, respectively. The leave-one-out cross validation demonstrated that for all models, 25% had $\leq$0.5 AU, 50% had $\leq$1.0 AU, and 75% had $\leq$1.7 AU distance error. The maximum observed error was 6.3 AU, with each model producing errors of $>$6.0 AU (Fig. 1).

Mapping antigenic predictions onto phylogenetic trees. Four trees were built with sequences genetically similar to four selected test antigens (Fig. 2). Trees were annotated with an amino acid motif based on positions 145, 155, 156, 158, 159, and 189, as these sites have been found to have a disproportionate effect on the observed

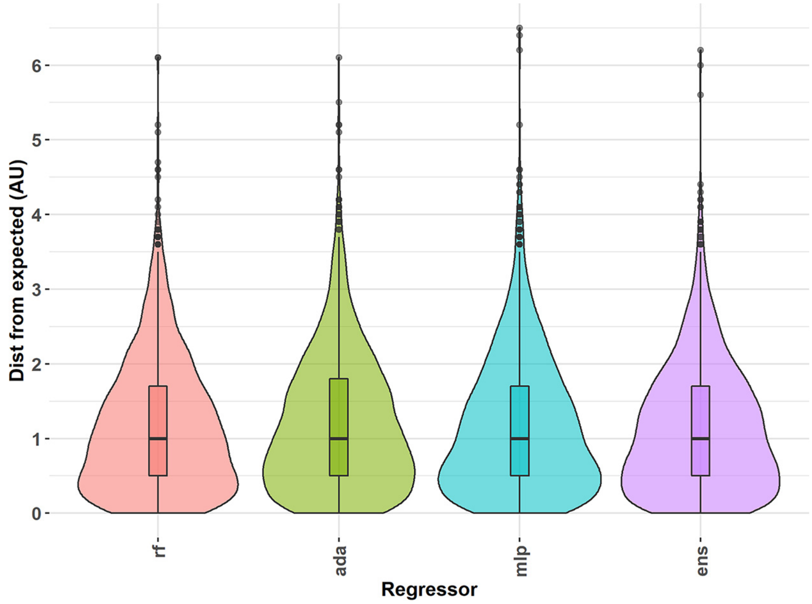

FIG 1 Distribution of errors calculated for the predicted antigenic distance compared to actual antigenic distance as predicted by machine learning models and hemagglutination inhibition assays, respectively. Three regression models were used to predict distances from empirically determined antigens using hemagglutination inhibition titers in a leave-one-out approach: random forest regression (rf), AdaBoost decision tree regression (ada), and multilayer perceptron (mlp) regression. All three predictions were combined into an ensemble (ens) to prevent overfitting and to minimize errant predictions by averaging across predictions from all models. Approximately 25% of the data have 0.5 antigenic units (AU) of error or less, and 50% of the data have 1 AU of error or less, with 75% of the data having less than 2 AU of error. Maximum error for outliers exceeded 6 AU.

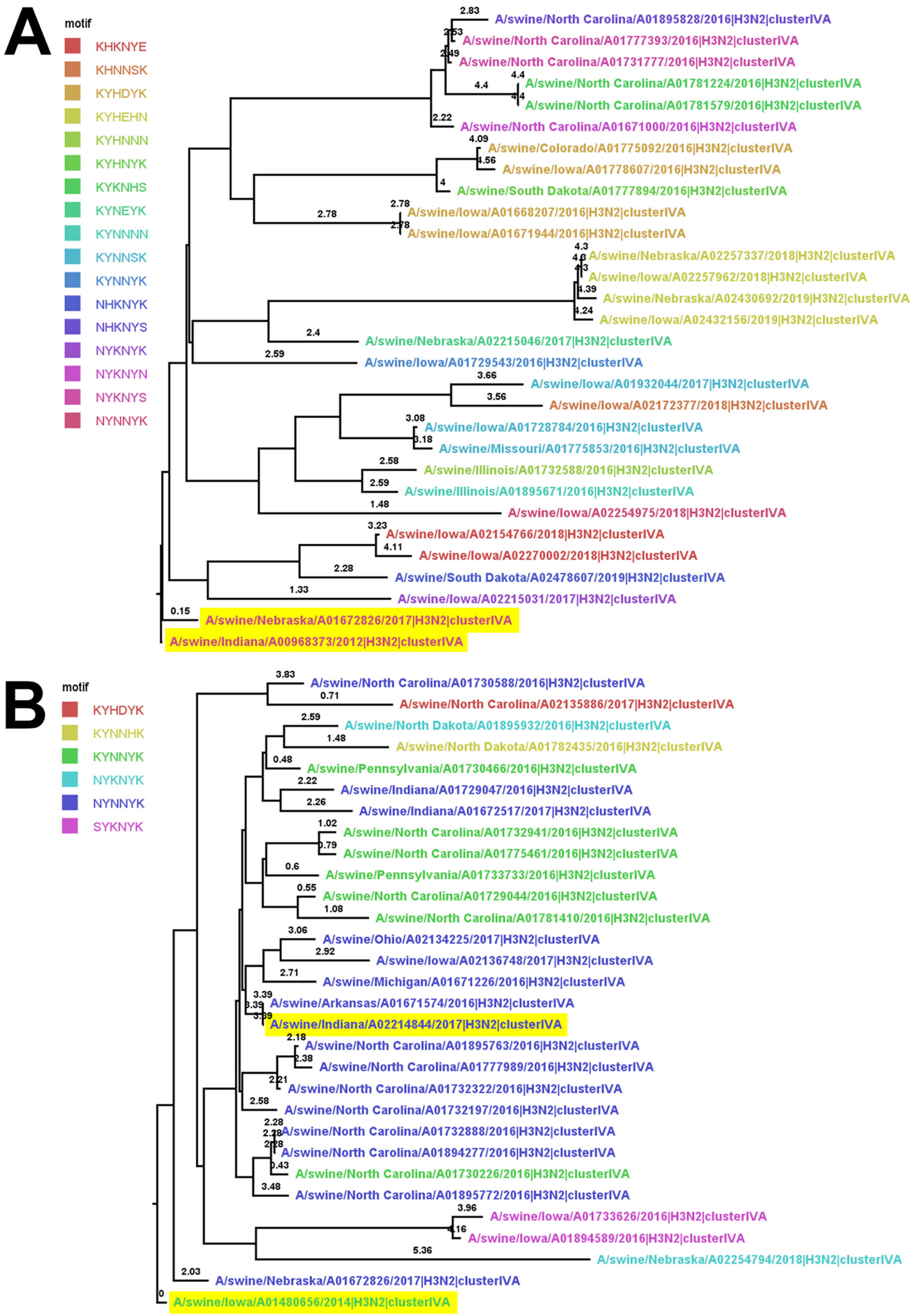
A
motif
KHKNYE
KHNNSK
KYHDYK
KYHEHN
KYHNNN
KYHNYK
KYKNHS
KYNEYK
KYNNNN
KYNNSK
KYNNYK
NHKNYK
NHKNYS
NYKNYK
NYKNYN
NYKNYS
NYNNYK
2.83
A/swine/North Carolina/A01895828/2016|H3N2|clusterIVA
2.53
A/swine/North Carolina/A01777393/2016|H3N2|clusterIVA
2.49
A/swine/North Carolina/A01731777/2016|H3N2|clusterIVA
4.4
4.4
A/swine/North Carolina/A01781224/2016|H3N2|clusterIVA
4.4
A/swine/North Carolina/A01781579/2016|H3N2|clusterIVA
2.22
A/swine/North Carolina/A01671000/2016|H3N2|clusterIVA
4.09
A/swine/Colorado/A01775092/2016|H3N2|clusterIVA
4.56
A/swine/Iowa/A01778607/2016|H3N2|clusterIVA
4
A/swine/South Dakota/A01777894/2016|H3N2|clusterIVA
2.78
2.78
A/swine/Iowa/A01668207/2016|H3N2|clusterIVA
2.78
A/swine/Iowa/A01671944/2016|H3N2|clusterIVA
4.3
4.3
A/swine/Nebraska/A02257337/2018|H3N2|clusterIVA
4.3
A/swine/Iowa/A02257962/2018|H3N2|clusterIVA
4.39
A/swine/Nebraska/A02430692/2019|H3N2|clusterIVA
4.24
A/swine/Iowa/A02432156/2019|H3N2|clusterIVA
2.4
A/swine/Nebraska/A02215046/2017|H3N2|clusterIVA
2.59
A/swine/Iowa/A01729543/2016|H3N2|clusterIVA
3.66
A/swine/Iowa/A01932044/2017|H3N2|clusterIVA
3.56
A/swine/Iowa/A02172377/2018|H3N2|clusterIVA
3.08
A/swine/Iowa/A01728784/2016|H3N2|clusterIVA
3.18
A/swine/Missouri/A01775853/2016|H3N2|clusterIVA
2.58
A/swine/Illinois/A01732588/2016|H3N2|clusterIVA
2.59
A/swine/Illinois/A01895671/2016|H3N2|clusterIVA
1.48
A/swine/Iowa/A02254975/2018|H3N2|clusterIVA
3.23
A/swine/Iowa/A02154766/2018|H3N2|clusterIVA
4.11
A/swine/Iowa/A02270002/2018|H3N2|clusterIVA
2.28
A/swine/South Dakota/A02478607/2019|H3N2|clusterIVA
1.33
A/swine/Iowa/A02215031/2017|H3N2|clusterIVA
0.15
A/swine/Nebraska/A01672826/2017|H3N2|clusterIVA
A/swine/Indiana/A00968373/2012|H3N2|clusterIVA
B
motif
KYHDYK
KYNNHK
KYNNYK
NYKNYK
NYNNYK
SYKNYK
3.83
A/swine/North Carolina/A01730588/2016|H3N2|clusterIVA
0.71
A/swine/North Carolina/A02135886/2017|H3N2|clusterIVA
2.59
A/swine/North Dakota/A01895932/2016|H3N2|clusterIVA
1.48
A/swine/North Dakota/A01782435/2016|H3N2|clusterIVA
0.48
A/swine/Pennsylvania/A01730466/2016|H3N2|clusterIVA
2.22
A/swine/Indiana/A01729047/2016|H3N2|clusterIVA
2.26
A/swine/Indiana/A01672517/2017|H3N2|clusterIVA
1.02
A/swine/North Carolina/A01732941/2016|H3N2|clusterIVA
0.79
A/swine/North Carolina/A01775461/2016|H3N2|clusterIVA
0.6
A/swine/Pennsylvania/A01733733/2016|H3N2|clusterIVA
0.55
A/swine/North Carolina/A01729044/2016|H3N2|clusterIVA
1.08
A/swine/North Carolina/A01781410/2016|H3N2|clusterIVA
3.06
A/swine/Ohio/A02134225/2017|H3N2|clusterIVA
2.92
A/swine/Iowa/A02136748/2017|H3N2|clusterIVA
2.71
A/swine/Michigan/A01671226/2016|H3N2|clusterIVA
3.39
3.39
A/swine/Arkansas/A01671574/2016|H3N2|clusterIVA
3.39
A/swine/Indiana/A02214844/2017|H3N2|clusterIVA
2.18
A/swine/North Carolina/A01895763/2016|H3N2|clusterIVA
2.38
A/swine/North Carolina/A01777989/2016|H3N2|clusterIVA
2.21
A/swine/North Carolina/A01732322/2016|H3N2|clusterIVA
2.58
A/swine/North Carolina/A01732197/2016|H3N2|clusterIVA
2.28
2.28
A/swine/North Carolina/A01732888/2016|H3N2|clusterIVA
2.28
A/swine/North Carolina/A01894277/2016|H3N2|clusterIVA
2.43
A/swine/North Carolina/A01730226/2016|H3N2|clusterIVA
3.48
A/swine/North Carolina/A01895772/2016|H3N2|clusterIVA
3.96
A/swine/Iowa/A01733626/2016|H3N2|clusterIVA
4.16
A/swine/Iowa/A01894589/2016|H3N2|clusterIVA
5.36
A/swine/Nebraska/A02254794/2018|H3N2|clusterIVA
2.03
A/swine/Nebraska/A01672826/2017|H3N2|clusterIVA
0
A/swine/Iowa/A01480656/2014|H3N2|clusterIVA

FIG 2 (Continued).

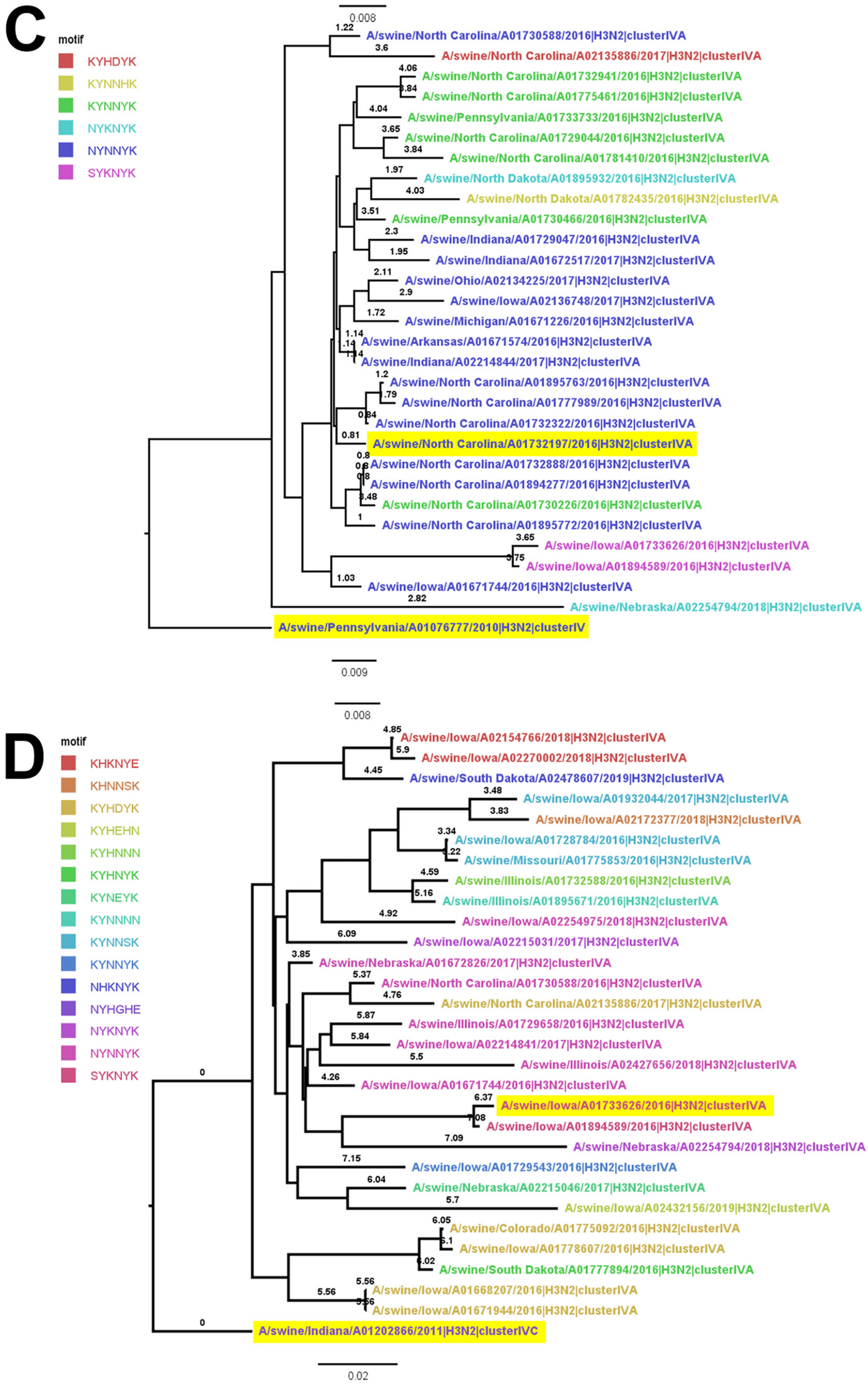
C
motif
KYHDYK
KYNNHK
KYNNYK
NYKNYK
NYNNYK
SYKNYK
0.008
1.22
A/swine/North Carolina/A01730588/2016|H3N2|clusterIVA
3.6
A/swine/North Carolina/A02135886/2017|H3N2|clusterIVA
4.06
A/swine/North Carolina/A01732941/2016|H3N2|clusterIVA
3.84
A/swine/North Carolina/A01775461/2016|H3N2|clusterIVA
4.04
A/swine/Pennsylvania/A01733733/2016|H3N2|clusterIVA
3.65
A/swine/North Carolina/A01729044/2016|H3N2|clusterIVA
3.84
A/swine/North Carolina/A01781410/2016|H3N2|clusterIVA
1.97
A/swine/North Dakota/A01895932/2016|H3N2|clusterIVA
4.03
A/swine/North Dakota/A01782435/2016|H3N2|clusterIVA
3.51
A/swine/Pennsylvania/A01730466/2016|H3N2|clusterIVA
2.3
A/swine/Indiana/A01729047/2016|H3N2|clusterIVA
1.95
A/swine/Indiana/A01672517/2017|H3N2|clusterIVA
2.11
A/swine/Ohio/A02134225/2017|H3N2|clusterIVA
2.9
A/swine/Iowa/A02136748/2017|H3N2|clusterIVA
1.72
A/swine/Michigan/A01671226/2016|H3N2|clusterIVA
1.14
1.14
A/swine/Arkansas/A01671574/2016|H3N2|clusterIVA
1.14
A/swine/Indiana/A02214844/2017|H3N2|clusterIVA
1.2
A/swine/North Carolina/A01895763/2016|H3N2|clusterIVA
1.79
A/swine/North Carolina/A01777989/2016|H3N2|clusterIVA
0.84
A/swine/North Carolina/A01732322/2016|H3N2|clusterIVA
0.81
A/swine/North Carolina/A01732197/2016|H3N2|clusterIVA
0.8
0.8
A/swine/North Carolina/A01732888/2016|H3N2|clusterIVA
0.8
A/swine/North Carolina/A01894277/2016|H3N2|clusterIVA
3.48
A/swine/North Carolina/A01730226/2016|H3N2|clusterIVA
1
A/swine/North Carolina/A01895772/2016|H3N2|clusterIVA
3.65
A/swine/Iowa/A01733626/2016|H3N2|clusterIVA
3.75
A/swine/Iowa/A01894589/2016|H3N2|clusterIVA
1.03
A/swine/Iowa/A01671744/2016|H3N2|clusterIVA
2.82
A/swine/Nebraska/A02254794/2018|H3N2|clusterIVA
A/swine/Pennsylvania/A01076777/2010|H3N2|clusterIV
0.009
D
motif
KHKNYE
KHNNSK
KYHDYK
KYHEHN
KYHNNN
KYHNYK
KYNEYK
KYNNNN
KYNNSK
KYNNYK
NHKNYK
NYHGHE
NYKNYK
NYNNYK
SYKNYK
0.008
4.85
A/swine/Iowa/A02154766/2018|H3N2|clusterIVA
5.9
A/swine/Iowa/A02270002/2018|H3N2|clusterIVA
4.45
A/swine/South Dakota/A02478607/2019|H3N2|clusterIVA
3.48
A/swine/Iowa/A01932044/2017|H3N2|clusterIVA
3.83
A/swine/Iowa/A02172377/2018|H3N2|clusterIVA
3.34
A/swine/Iowa/A01728784/2016|H3N2|clusterIVA
3.22
A/swine/Missouri/A01775853/2016|H3N2|clusterIVA
4.59
A/swine/Illinois/A01732588/2016|H3N2|clusterIVA
5.16
A/swine/Illinois/A01895671/2016|H3N2|clusterIVA
4.92
A/swine/Iowa/A02254975/2018|H3N2|clusterIVA
6.09
A/swine/Iowa/A02215031/2017|H3N2|clusterIVA
3.85
A/swine/Nebraska/A01672826/2017|H3N2|clusterIVA
5.37
A/swine/North Carolina/A01730588/2016|H3N2|clusterIVA
4.76
A/swine/North Carolina/A02135886/2017|H3N2|clusterIVA
5.87
A/swine/Illinois/A01729658/2016|H3N2|clusterIVA
5.84
A/swine/Iowa/A02214841/2017|H3N2|clusterIVA
5.5
A/swine/Illinois/A02427656/2018|H3N2|clusterIVA
4.26
A/swine/Iowa/A01671744/2016|H3N2|clusterIVA
6.37
A/swine/Iowa/A01733626/2016|H3N2|clusterIVA
7.08
A/swine/Iowa/A01894589/2016|H3N2|clusterIVA
7.09
A/swine/Nebraska/A02254794/2018|H3N2|clusterIVA
7.15
A/swine/Iowa/A01729543/2016|H3N2|clusterIVA
6.04
A/swine/Nebraska/A02215046/2017|H3N2|clusterIVA
5.7
A/swine/Iowa/A02432156/2019|H3N2|clusterIVA
6.05
A/swine/Colorado/A01775092/2016|H3N2|clusterIVA
6.1
A/swine/Iowa/A01778607/2016|H3N2|clusterIVA
6.02
A/swine/South Dakota/A01777894/2016|H3N2|clusterIVA
5.56
5.56
A/swine/Iowa/A01668207/2016|H3N2|clusterIVA
5.56
A/swine/Iowa/A01671944/2016|H3N2|clusterIVA
0
0
A/swine/Indiana/A01202866/2011|H3N2|clusterIVC
0.02

TABLE 2 Hemagglutination inhibition titers representing the homologous reference strain titer and heterologous test antigen titer[a]

	Titer for indicated serum strain							
Test antigen	A/swine/Indiana/A00968373/2012	A/swine/Indiana/A00968373/2012	A/swine/Iowa/A01480656/2014	A/swine/Iowa/A01480656/2014	A/swine/Pennsylvania/A01076777/2010	A/swine/Pennsylvania/A01076777/2010	A/swine/Indiana/A01202866/2011	A/swine/Indiana/A01202866/2011
A/swine/Indiana/A00968373/2012	**640**	**2,560**						
A/swine/Nebraska/A01672826/2017	1,280	2,560						
A/swine/Iowa/A01480656/2014			**1,280**	**2,560**				
A/swine/Indiana/A02214844/2017			160	80				
A/swine/Pennsylvania/A01076777/2010					**2,560**	**640**		
A/swine/North Carolina/A01732197/2016					320	160		
A/swine/Indiana/A01202866/2011							**5,120**	**5,120**
A/swine/Iowa/A01733626/2016							40	80
Log_2 difference	−1	0	3	5	3	2	7	6

[a] Each titer was determined in duplicate, with homologous and heterologous titers determined in parallel. Boldface indicates homologous titer.

antigenic phenotype in both human and swine H3 (14). The antigenic motifs of test antigen A/swine/Nebraska/A01672826/2017 and reference antiserum A/swine/Indiana/A00968373/2012 match, both being NYNNYK (Fig. 2A). The antigenic motif of test antigen A/swine/Indiana/A02214844/2017 was NYNNYK, while reference antiserum A/swine/Iowa/A01480656/2014's motif was KYNNYK, differing at position 145 (Fig. 2B). The antigenic motifs of test antigen A/swine/North Carolina/A01732197/2016 and reference antiserum A/swine/Pennsylvania/A01076777/2010 match, both being NYNNYK (Fig. 2C). The antigenic motif of test antigen A/swine/Iowa/A01733626/2016 was SYKNYK, while reference antiserum A/swine/Indiana/A01202866/2011's motif was NYHGHE, differing at positions 145, 156, 158, 159, and 189 (Fig. 2D).

Empirical validation of the predicted antigenic distance predictions. The predicted ensemble distances of the four selected test antigens were validated via HI assay (Table 2). Test antigen A/swine/Nebraska/A01672826/2017 was predicted to be 0.15 AU from reference strain A/swine/Indiana/A00968373/2012, with 99.4% amino acid identity shared between the HA1 segments of the HA (Table 3). Both the reference and test antigens were from the H3 cluster IVA clade (Fig. 2A), and this pairing represents a near identity and near antigenic distance prediction. The amino acid differences between the reference strain and the test antigen were at M10T and R208I (Table 3). The HI assay demonstrated that the antigenic distance between the refer-

FIG 2 Phylogenetic trees of test antigens rooted to their reference strain. (A) Phylogenetic tree of test antigen A/swine/Nebraska/A01672826/2017 and reference strain A/swine/Indiana/A00968373/2012, representing a near predicted antigenic distance prediction (0.15 AU) for two strains of near amino acid identity (99.4%). (B) Phylogenetic tree of test antigen A/swine/Indiana/A02214844/2017 and reference strain A/swine/Iowa/A01480656/2014, representing a far predicted antigenic distance prediction (3.39) for two strains of near amino acid identity (98.5%). (C) Phylogenetic tree of test antigen A/swine/North Carolina/A01732197/2016 and reference strain A/swine/Pennsylvania/A01076777/2010, representing a near predicted antigenic distance prediction (0.81) for two strains of far amino acid identity (94.2%). (D) Phylogenetic tree of test antigen A/swine/Iowa/A01733626/2016 and reference strain A/swine/Indiana/A01202866/2011, representing a far predicted antigenic distance prediction (6.37) for two strains of far amino acid identity (91.2%). Branches of the phylogenetic tree were annotated with the predicted antigenic distance from the ensemble regression model (both test antigen and reference strain are highlighted). Each tree is pruned to 30 sequences. Influenza virus strains are colored by the antigenic motif formed by amino acid positions 145, 155, 156, 158, 159, and 189; these positions, located near the ligand binding site of the hemagglutinin protein, have been noted to affect the antigenic interactions of the protein.

TABLE 3 Amino acid mutations detected between test antigen and reference strains used for the model validation

Test antigen	Reference strain	Amino acid changes
A/swine/Nebraska/A01672826/2017	A/swine/Indiana/A00968373/2012	M10T, R208I
A/swine/Indiana/A02214844/2017	A/swine/Iowa/A01480656/2014	G49S, E83K, V112I, K145N, S289P
A/swine/North Carolina/A01732197/2016	A/swine/Pennsylvania/A01076777/2010	T10M, E83K, V106A, S107T, V112I, T117N, N124S, K142S, A163E, L164Q, M168V, N173K, I196V, T203I, P273H, G275D, N276E, K278N, R299K, V304A
A/swine/Iowa/A01733626/2016	A/swine/Indiana/A01202866/2011	I29L, G50R, E83K, S107T, T117N, S124N, A131D, D133G, R137N, S138T, R140K, G144V, N145S, H156K, G158N, H159Y, A163E, L164Q, T167A, N173K, E189K, S193N, V196A, I203V, R220V, R269K, S273H, N276E, R299K

ence strain antiserum and test antigen was 0.5 AU (Table 4), with an error between the predicted distance and the empirical distance of 0.35 AU.

Test antigen A/swine/Indiana/A02214844/2017 was predicted to be 3.39 AU from reference strain A/swine/Iowa/A01480656/2014, with 98.5% amino acid identity shared between the HA1 segments. Both the reference strain and test antigens are from the H3 cluster IVA clade (Fig. 2B), and this pairing represents near identity but far antigenic distance prediction. There were 5 amino acid differences between the reference strain and test antigen (Table 3). The HI assay found a distance of 4.0 AU between the test antigen and reference antiserum and an error of 0.61 AU between empirical and predicted distances (Table 4).

Test antigen A/swine/North Carolina/A01732197/2016 was predicted to be 0.81 AU from reference strain A/swine/Pennsylvania/A01076777/2010, with 93.9% amino acid identity shared between the HA1 segments. The test antigen was selected from the H3 cluster IVA clade, and the reference strain was selected from the H3 cluster IV clade (Fig. 2C); this pair represents a far identity, but the antigen and reference strain were predicted to be antigenically similar. There were 20 amino acid differences between the reference strain and test antigen (Table 3). The HI assay demonstrated an average antigenic distance between reference antiserum and test antigen of 2.5 AU, with a prediction error of 1.69 AU (Table 4).

Test antigen A/swine/Iowa/A01733626/2016 was predicted to be 6.37 AU from reference strain A/swine/Indiana/A01202866/2011, with 91.2% amino acid identity shared between the HA1 segments. The test antigen is from the H3 cluster IVA clade of virus, and the reference strain is from the H3 cluster IVC clade (Fig. 2D). This pairing represents a far identity and far predicted antigenic distance prediction. There were 29 amino acid differences between the reference strain and the test strain (Table 3). The HI assay demonstrated 6.5 AU between test antigen and reference antiserum, giving an error of 0.13 AU between empirical and predicted distances (Table 4).

Ranking of predictor features. Random forest regression, one of the regressors composing the ensemble model, ranks user-selected features by a metric of importance, calculated by the decrease in the node variance per tree and normalized across the forest for a single model run so that the sum of importance scores is equal to 1 (22) (Table S2). The highest-ranking features were stable across runs, as they had a consistent decrease in

TABLE 4 Predicted and measured antigenic distances between test antigens and reference strain antisera using the model to calculate the predicted distance and HI titers to calculate the empirical distance in antigenic units

Test antigen	Reference antiserum	Test antigen motif	Amino acid identity (%)	Predicted distance (AU)	HI distance (AU)	Error (AU)[a]
A/swine/Nebraska/A01672826/2017	A/swine/Indiana/A00968373/2012	NYNNYK	99.4 (near)	0.15 (near)	0.5	0.35
A/swine/Indiana/A02214844/2017	A/swine/Iowa/A01480656/2014	NYNNYK	98.5 (near)	3.39 (far)	4.0	0.61
A/swine/North Carolina/A01732197/2016	A/swine/Pennsylvania/A01076777/2010	NYNNYK	93.9 (far)	0.81 (near)	2.5	1.69
A/swine/Iowa/A01733626/2016	A/swine/Indiana/A01202866/2011	SYKNYK	91.2 (far)	6.37 (far)	6.5	0.13

[a] The error was calculated by subtracting the absolute value of the predicted distance from the empirical distance.

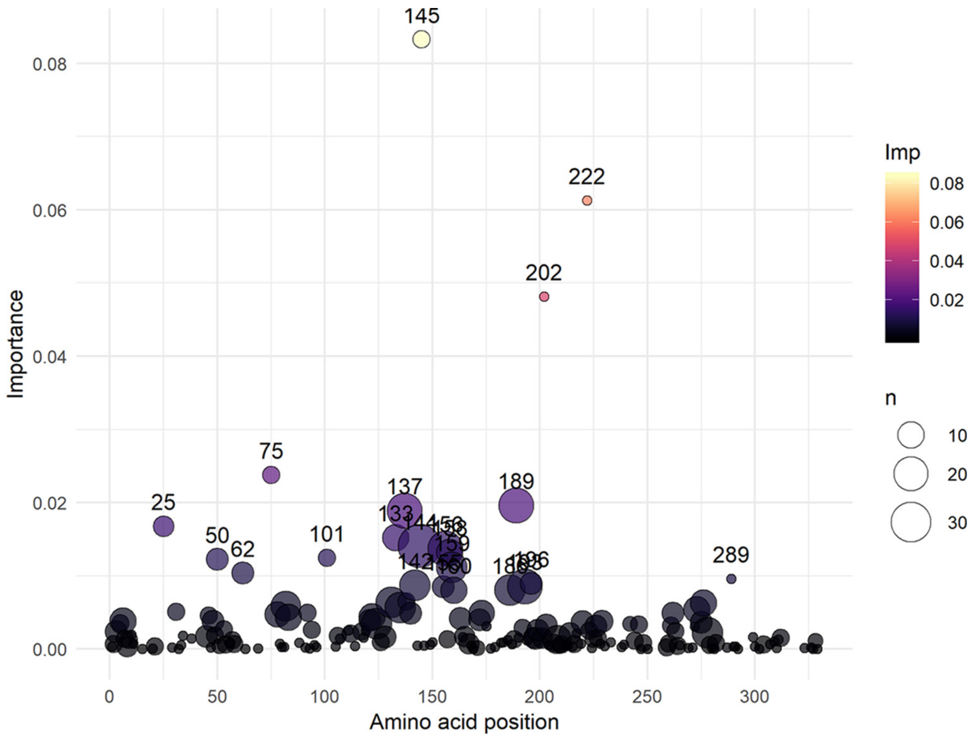

FIG 3 Rank of amino acid location importance by the cumulative summation of importance per site mutation as determined by random forest regression. Amino acid position using H3 numbering is reported on the *x* axis. The importance for each site-specific mutation is summed per site and displayed on the *y* axis using a color scale. The sum of importance is scaled to 1 and is unitless. The size of the circle is relative to the number of mutations observed in the training set per site. Identity was the highest-ranking feature, with an importance of 0.312, but is not displayed on the graph. The top 10 amino acid transition features in order of importance are K145N, I202V, R222W, H75Q, R137Y, D101Y, E62K, I25L, P289S, and D133N. The top 10 amino acid sites in order of cumulative importance are 145, 202, 222, 75, 189, 137, 144, 133, 156, and 101.

their average variance, although these metrics were susceptible to starting conditions (data provided at https://github.com/flu-crew/antigenic-prediction). The most important feature in predicting the antigenic distance between two strains was amino acid identity within the HA1, accounting for 31.4% of the importance score. Transitions between K and N at position 145 accounted for 8.1% of the model's importance score, and this change was ranked as the most important amino acid mutation. However, transitions between K and S and N and S at the same position 145 received a lower ranking in the model's importance score (totaling 0.2% importance cumulatively), demonstrating that the context of the positional mutation is important. Features I202V and R222W (representing bidirectional mutations) accounted for 5.4% and 5.2% of the importance score, respectively. The remainder of the features in the models accounted for less than 3% of the model on an individual basis (Fig. 3; see Table S2 in the supplemental material), with the next 10 bidirectional mutations in order of importance being H75Q, R137Y, D101Y, E62K, I25L, P289S, D133N, E189K, K92T, and H159Y (Fig. 3). Projecting the cumulative importance of each amino acid position on an H3 crystal structure indicated that position 145, the most important position in the model, is located in the groove of the active site (Fig. 4). Other sites of higher importance in the model were more likely to be observed on the solvent-facing side of the trimer. Amino acid position 202 was an exception, as it was ranked as having high importance but was located on the inside of the trimer.

Of the 728 features included in the model, amino acid identity and the sum of the top 10 amino acid mutation features of the model accounted for 58.3% of the model's importance. The top 100 features, including percent identity within the HA1 and amino acid mutations, accounted for 83% of the calculated importance. The top

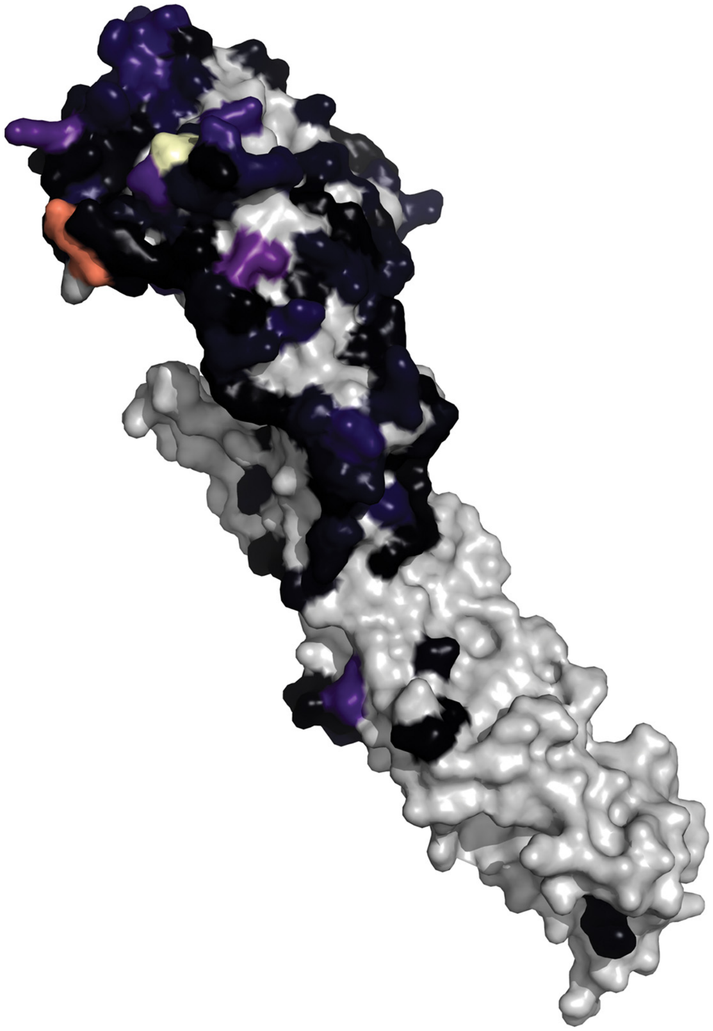

FIG 4 Projection of feature importance on a monomer of the A/Victoria/361/2011 hemagglutinin (HA) protein (RCSB 4O5N). The importance for each site-specific mutation is summed per site and projected onto the hemagglutinin protein model of the human H3. Higher color intensity represents a larger calculated importance. Positions with no data are colored gray.

253 amino acid mutation features and percent identity accounted for 95% of the calculated importance. The model required 397 features along with percent identity to account for 99% of the calculated importance.

DISCUSSION

In this study, a model was developed to computationally estimate antigenic distances between different IAVs in swine based on amino acid sequence using nonlinear machine learning methods. The method leverages data that were generated from previously characterized IAV strains in swine to train regression models. After *in silico* validation, the models were used to predict the antigenic distance between paired IAV strains based on amino acid identity and mutations present between each strain. The antigenic predictions were experimentally confirmed by comparing the distances between homologous and heterologous hemagglutination inhibition (HI) titers. Predicting antigenic distances from genetic sequence data can identify strains that require further antigenic characterization, reduce the number of HI assays required to describe circulating antigenic diversity, and aid in the selection of candidate strains for vaccines when genetic diversity surveilled in the field does not have an adequate antigenic match in current vaccine formulations.

This work adds to a growing body of literature that aims to quantitatively predict antigenic phenotypes of IAV from the sequence without requiring HI titers for each IAV strain (19, 23–26). To the best of our knowledge, earlier approaches to calculate antigenic distances between IAV strains were trained and tested on human IAV strains, where the HA genes are characterized by phylogenetic trees with a single thick trunk with short interspersed branches with far less cocirculating genetic diversity (27–29). Compared to IAVs circulating in humans, HA gene phylogenetic trees from endemic IAVs in swine demonstrate multiple genetic clades within the same subtype that are derived from multiple human-to-swine spillover events across the last 100 years (7, 30). The large genetic diversity of strains coevolving within the swine population has resulted in a similarly large breadth of antigenic diversity and evolution. Consequently, a broad range of HI assays including many genetically different IAVs are needed to assess the antigenic diversity of IAVs circulating within swine. The scale of these studies has been difficult, and there is a sparsity of antigenic characterization of IAV in swine, frequently with large gaps of time between characterizations. This has the unfortunate consequence of potentially misrepresenting the antigenic diversity of swine IAVs and can make it difficult to improve our understanding of antigenic evolution of IAV in swine (19, 26, 31).

We experimentally validated our model using four test antigens, with the empirical data demonstrating that predictions generally had an error of less than 1 AU. These four strains were selected to represent the full spectrum of observed diversity within the H3 cluster IV genetic clade. Our model performed very well on sequences with high sequence identity that were predicted to be antigenically similar (near identity/near distance = 0.5 AU, with 0.35 AU error) (Tables 2 and 4). Similarly, the model performed well when making predictions on sequences that were genetically similar but predicted to be antigenically distinct (near identity/far distance = 4 AU, with 0.61 AU error) (Tables 2 and 4) and those that were very genetically different and predicted to be antigenically distinct (far identity/far distance = 6.37 AU, with 0.13 AU error) (Tables 2 and 4). On sequences that were genetically dissimilar but were predicted to be antigenically similar, the model had a nonnegligible error (1.69 AU); however, the ensemble prediction was able to discern that these two strains were more antigenically similar than would be predicted based on sequence similarity alone. The large error in this prediction, despite all features being accounted for in the model (see Table S2 in the supplemental material), suggests limitations in our approach. We parameterized the model with empirical data, and sequences that fit the "far identity/near antigenic distance" are very sparse in the training set, resulting in a higher prediction error. As new empirical data are generated, they can be used to refine and improve the model. This point is also valid for the "near identity/far antigenic distance" predictions that were parameterized by a small number of empirical observations. It should be noted that the HI assay is a discrete measure whereas the prediction is continuous, and thus an error of less than 1 AU is not biologically meaningful. Additionally, because of the discrete nature of the HI assay, a 0.5 AU error is negligible, as the true antigenic distance is somewhere between 0 and 1 AU. Consequently, our approach, which was developed using a relatively small empirical data set of IAV in swine, made predictions that are useful in biological applications.

An additional benefit of machine learning methods is that they can assign an importance score to the position and context of amino acid mutations, allowing biological interpretation. This importance score is calculated by the decrease in the node variance after fitting the random forest model. While sequence amino acid difference had the highest importance score, further assessment of the model revealed that both the position and the context of the amino acid mutation contributed to the observed antigenic phenotype. An example of this dynamic was H3 HA position 145, where a mutation between K and N bidirectionally was ranked as the most important amino acid mutation feature. Other observed mutations at position 145 between K and S and N and S were less important, matching the biological nuances that have been observed with empirical testing and other computational predictions (15, 24). Earlier literature has suggested that conservation of biochemical properties of the amino acid mutation may also have some effect on the observed antigenic change (15, 19). Sites other than these were identified as important in determining phenotype and were located on the solvent-exposed surface of the HA protein and in antibody epitopes (Fig. 4) (32, 33). The positions in our model demonstrated overlap with those of a human IAV machine learning algorithm (23), the joint random forest regression (JRFR) algorithm (positions 62, 121, 131, 133, 135, 137, 142, 144, 145, 155, 156, 158, 159, 172, 173, 189, 193, 196, and 276) (Table S2), but the relative importance of the predictor features varied between this model and ours. Specifically, position 189 was the most important site in human H3 with ferret antisera, whereas our model identified position 145 as the most important position in swine H3 with swine sera (23). These differences are likely to be reflective of host-specific interactions, and there is evidence that the source of antisera may impact HI results (34). Additionally, our importance ranking demonstrated that a relatively small number of sites had a disproportionate importance for the phenotype (Fig. 3). Consequently, these data suggest that incorporating the identity of amino acid mutation alongside sequence homology will help improve vaccine antigen selection, as this likely has a critical influence on antigen-antibody interactions.

There are other *in silico* approaches that link genetic sequence data to antigenic phenotype. Using 10-fold cross validation, our ensemble model had a higher RMSE (1.21 AU) than JRFR, a random forest-based model that consistently has an RMSE of $<$1.0 (23). Similarly, the linear mixed-effects model employed by Harvey had very strong performance (mean absolute error, 0.75 U) (26). However, a direct comparison between these and similar methods used in human IAV with our approach is difficult because of the major differences between extensive training data sets and our own and the observed genetic diversity of swine IAVs with multiple cocirculating lineages (26). Our approach does have utility, as the robust leave-one-out cross validation demonstrated that 54% of the predictions made with the ensemble model were at or below 1 AU of error, and 86% were below 2 AU of error (a distance of $<$2 AU is frequently used to indicate biological equivalence), and we were able to experimentally validate our *in silico* predictions with strains that represented the full spectrum of genetic diversity in H3 cluster IV swine IAVs.

Our ensemble of nonlinear regression methods was chosen due to a nonlinear relationship that is not strictly additive between amino acid changes and antigenic phenotype. The nonlinear regression techniques used are robust against collinearity, and the tree methods have the benefit of ranking the contribution of each feature to the predictive power of the model, designated through an importance score (22, 35). These data can subsequently be used to inform *in vitro* or *in vivo* studies that determine molecular features associated with antibody recognition and drift (14). Several earlier methods implement linear regression, despite the relationship between amino acid mutation and antigenic phenotype being nonlinear and not strictly additive (19, 25). Linear models can mitigate issues of collinearity by implementing approaches such as ridge regression in antigen bridges (24) or lasso regression used by Nextstrain (19, 31), but these approaches may result in models that are more difficult to interpret biologically. Consequently, our empirically validated models, although not as computationally accurate, performed in a biologically meaningful manner and were also able to identify the top 10 features accounting for 58.3% of the antigenic phenotype (253 features were needed to account for 95% importance). These data have now generated explicit predictions on when specific mutations in the HA gene may result in antigenic drift and reduce vaccine efficacy.

Our experimental validation using test antigen and reference strains demonstrated that this approach can be used to determine antigenic differences between IAVs without requiring extensive HI testing in laboratories. It is currently impractical to antigenically characterize all strains of IAV isolated from swine, and our work shows that antigenic phenotype can be reasonably predicted from genetic sequence. The performance of our approach was sufficient even though it was parameterized with a limited empirical data set; it is feasible that prediction can be improved as more empirical data are made available. Due to multiple introductions of IAV into swine from human and avian sources, the genetic diversity of IAV in swine exceeds what is observed for human IAV strains (11, 30, 36). The genetic diversity of IAV in swine is also confounded by transportation patterns that move regional IAV strains with swine to new geographic locations, where additional antigenic drift and reassortment with endemic strains may occur (37, 38). Consequently, this method can aid in vaccine design efforts for IAV in swine, which currently do not have an integrated and comprehensive system such as the World Health Organization's (WHO) global influenza surveillance program for IAV in humans (39). Providing accurate methods such as ours that predict antigenic distances of IAV in swine increases the ability of swine producers and veterinarians to make informed decisions regarding vaccine antigens to help maintain swine herd health.

MATERIALS AND METHODS

Swine IAV H3 antigenic reference data set. The antigenic properties of two influenza viruses can be quantitatively compared using a hemagglutination inhibition (HI) assay. The assay is based on the ability of the hemagglutinin to agglutinate red blood cells, which express sialic acid on their cell surface (40, 41). The HI antibodies raised against a homologous IAV can block the agglutination of red blood cells, even at low concentrations. Genetically different viruses often need a higher concentration of HI antibodies to prevent agglutination than the homologous titer. Comparing the antigenic distances between two viruses is calculated by distance $D_{ij} = \log_2(H_{jj}) - \log_2(H_{ij})$, representing a 2-fold loss in HI antibody cross-reactivity between the homologous and heterologous HI antibody titers (42) (H_{ij} represents the titer between heterologous serum *i* and antigen *j*, and H_{jj} represents a homologous titer). These data have traditionally been used to generate pairwise antigenic distances between IAVs in swine that are then visualized using multidimensional scaling to form an antigenic map (9, 43, 44).

The HI titers were collected from prior swine H3 HA virus characterization studies that used HI assays (41, 45, 46). The HI titers from new IAVs selected as reference strains were collected at the time of the experiment to expand the data set by the use of methods described in earlier literature, totaling 128 reference antigens tested against 47 reference antisera in various combinations from combined experiments (40). Distances between available HI titers were calculated by subtracting the $\log_2$ of the heterologous titer from the $\log_2$ of the homologous titer (42). Distances corresponding to the same antigen-antiserum pair were calculated as the $\log_2$ of the geometric mean by the following equation:

$$\overline{D}_{ij} = \frac{\log_2\left(\frac{H_{jj_1} H_{jj_2}}{H_{ij_1} H_{ij_2}}\right)}{2}$$

Training and validation of machine learning regression models. Full-length HA amino acid sequences for each antigen represented in the data set were aligned using MAFFT v7.311 (47) and then trimmed to the HA1 domain (amino acids 1 to 328 using the H3 HA numbering with the signal peptide removed) for subsequent analyses. Percent amino acid difference (100% − amino acid identity) was calculated between each HA pair for all combinations of sequences. Specific amino acid substitutions were not weighted to minimize model assumptions, and prior research in human IAV has suggested that these approaches may add noise to analysis (23, 48). All observed site-specific amino acid substitutions in the reference data were identified and treated as bidirectional.

The regression model data were constructed with the antigenic distance calculated from the HI titer as the training value, with the percent amino acid difference as a continuous predictor feature and site-specific mutations as binary predictor features. Three different machine learning regression models were trained using scikit-learn (49): random forest, AdaBoost decision tree, and multilayer perceptron. For each regression model, hyperparameters were tuned using a random search optimization (see Table S1 in the supplemental material). A fourth regression model was created by averaging the three prior machine learning model predictors and is referred to as the ensemble model.

Data were split into 80% training and 20% testing data groups to calculate the Pearson correlation and root mean square error. Additionally, 10-fold cross validation was used to assess the root mean square error (Table 1). Given the sparsity of antigenic data available, a leave-one-out cross validation approach was employed to generate a distribution of prediction errors for each model (Fig. 1). Each antigen included in the training set ($n = 128$) was iteratively excluded from the training set, and distances were predicted by using each of the four regression models. The error was calculated as the absolute value of difference between the predicted distance and the empirical distance.

Mapping antigenic predictions onto phylogenetic trees. Maximum-likelihood phylogenetic trees were created to assess antigenic distance predictions of genetically similar sequences of the test antigen sequence compared to the reference sequence. Sequences were aligned using MAFFT v7.311 (47), and phylogenetic trees were inferred using FastTree v2.1.10 (50). Trees were annotated using FigTree v1.4.3 (51), with each tree rooted to a reference strain and sorted in ascending order relative to the inferred evolutionary relationship. Each tip within the tree was color coded based on the antigenic motif designated by H3 numbering of positions 145, 155, 156, 158, 159, and 189, as earlier work had identified these sites as significant for antigenic phenotype (15). Branches were annotated with the ensemble-predicted antigenic distance relative to the root. Trees were pruned to 30 leaves to facilitate viewing.

Determining the relative importance of genetic mutations. Random forest regression models provide a natural ranking system of feature importance (22, 35). The importance of each predictor feature was calculated by the decrease in the node variance after fitting the random forest model. The feature importance rankings for the random forest regression model were analyzed to assess the biological importance of observed mutations in the swine H3 antigenic reference data set. The significance of each amino acid position in the HA was determined by summing the mutation-based features grouped by the position they represented. The resultant significance of each amino acid was projected onto a protein model of a human H3 HA gene from strain A/Victoria/361/2011 obtained from the Research Collaboratory for Structural Bioinformatics (4O5N) (52).

Empirical validation of machine learning regression models. The H3 HA amino acid sequences of uncharacterized IAVs in swine submitted to NCBI GenBank from the Iowa State University Veterinary Diagnostic Lab from January 2016 to August 2018 were collected and clustered by phylogenetic clade (7, 11). The HA gene sequences were trimmed to the HA1 domain (positions 1 to 328 using H3 numbering with the signal peptide removed). The HA1 sequences were compared against all antigenically characterized sequences to calculate percent amino acid difference and to compare the presence or absence of site-specific amino acid mutations. Site-specific amino acid mutations absent from the training set were not considered in additional analyses. The antigenic distance from each uncharacterized HA gene to each reference antigen was predicted using the previously described four trained regression models.

A selection of four contemporary IAVs were selected as test antigens to be antigenically characterized with *in vitro* HI assays to validate the regression models by using their HA genes. We selected these HA genes from within the H3 cluster IVA genetic clade, since (i) this is a significant genetic clade that is frequently detected in diagnostic submissions to the Iowa State University Veterinary Diagnostic Lab (11), (ii) this genetic clade was responsible for more than 300 zoonotic infections from 2012 to present, and (iii) there was a significant amount of uncharacterized data for this clade within the last 2 years ($n = 299$ from 2018 to present, representing 8% of sequenced HA genes). Since the ensemble predictions demonstrated the least error in the analyses above, antigenic distances of 106 H3 cluster IVA viruses were predicted against a panel of 44 available antisera using this model. We selected four test antigen/antiserum prediction pairs within this genetic clade based on the following criteria: near amino acid sequence identity (≥98%) and near predicted ensemble antigenic distance measured in antigenic units (AU) (≤2 AU); near identity and far antigenic distance (≥3 AU); far identity (≤95%, ≥90%) and near antigenic distance (≤2 AU); or far identity (≤95%, ≥90%) and far antigenic distance (≥3 AU) (Fig. 2; Table 4).

The four selected antigen/antiserum pairs were tested in parallel with antigens homologous to the antisera via HI assay. HI assays were conducted as previously described (41), with empirical distances calculated by subtracting the $\log_2$ of the heterologous titer from the $\log_2$ of the homologous titer. Empirical distances were compared against predicted values by subtraction.

Data availability. Data and code used in this research are available in a GitHub repository (https://github.com/flu-crew/antigenic-prediction).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TABLE S1, PDF file, 0.04 MB.

TABLE S2, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We gratefully acknowledge pork producers, swine veterinarians, and laboratories for participating in the USDA Influenza A Virus in Swine Surveillance System and publicly sharing sequences in NCBI GenBank.

This work was supported by the following: the Iowa State University Presidential Interdisciplinary Research Initiative; the Iowa State University Veterinary Diagnostic Laboratory; the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS project number 5030-32000-120-00-D); a National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health interagency agreement associated with the Center of Research in Influenza Pathogenesis, an NIAID-funded Center of Excellence in Influenza Research and Surveillance (grant number HHSN272201400008C to A.L.V.); the USDA Agricultural Research Service Research Participation Program of the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the USDA Agricultural Research Service (contract number DE-AC05-06OR23100 to Z.W.A. and C.K.S.); the Department of Defense, Defense Advanced Research Projects Agency, Preventing Emerging Pathogenic Threats program (HR00112020034 to P.C.G, T.K.A., and A.L.V.); and the SCINet project of the USDA

Agricultural Research Service (ARS project number 0500-00093-001-00-D). Funding for open access charges was from the U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS project number 5030-32000-120-00-D).

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA, DOE, or ORISE. USDA is an equal opportunity provider and employer.

REFERENCES

1. Dykhuis- Haden C, Painter T, Fangman T, Holtkamp D. 2012. Assessing production parameters and economic impact of swine influenza, PRRS and Mycoplasma hyopneumoniae on finishing pigs in a large production system. Abstr Am Assoc Swine Veterinarians, Denver, CO.
2. Saitou N, Nei M. 1986. Polymorphism and evolution of influenza A virus genes. Mol Biol Evol 3:57–74. https://doi.org/10.1093/oxfordjournals.molbev.a040381.
3. Sandbulte MR, Spickler AR, Zaabel PK, Roth JA. 2015. Optimal use of vaccines for control of influenza A virus in swine. Vaccines (Basel) 3:22–73. https://doi.org/10.3390/vaccines3010022.
4. Vincent AL, Ciacci-Zanella JR, Lorusso A, Gauger PC, Zanella EL, Kehrli ME, Jr, Janke BH, Lager KM. 2010. Efficacy of inactivated swine influenza virus vaccines against the 2009 A/H1N1 influenza virus in pigs. Vaccine 28:2782–2787. https://doi.org/10.1016/j.vaccine.2010.01.049.
5. Van Reeth K, Labarque G, De Clercq S, Pensaert M. 2001. Efficacy of vaccination of pigs with different H1N1 swine influenza viruses using a recent challenge strain and different parameters of protection. Vaccine 19:4479–4486. https://doi.org/10.1016/S0264-410X(01)00206-7.
6. Vincent AL, Lager KM, Janke BH, Gramer MR, Richt JA. 2008. Failure of protection and enhanced pneumonia with a US H1N2 swine influenza virus in pigs vaccinated with an inactivated classical swine H1N1 vaccine. Vet Microbiol 126:310–323. https://doi.org/10.1016/j.vetmic.2007.07.011.
7. Anderson TK, Nelson MI, Kitikoon P, Swenson SL, Korslund JA, Vincent AL. 2013. Population dynamics of cocirculating swine influenza A viruses in the United States from 2009 to 2012. Influenza Other Respir Viruses 7 (Suppl 4):42–51. https://doi.org/10.1111/irv.12193.
8. Walia RR, Anderson TK, Vincent AL. 2019. Regional patterns of genetic diversity in swine influenza A viruses in the United States from 2010 to 2016. Influenza Other Respir Viruses 13:262–273. https://doi.org/10.1111/irv.12559.
9. Lewis NS, Russell CA, Langat P, Anderson TK, Berger K, Bielejec F, Burke DF, Dudas G, Fonville JM, Fouchier RA, Kellam P, Koel BF, Lemey P, Nguyen T, Nuansrichy B, Peiris JM, Saito T, Simon G, Skepner E, Takemae N, Consortium E, Webby RJ, Van Reeth K, Brookes SM, Larsen L, Watson SJ, Brown IH, Vincent AL, ESNIP3 Consortium. 2016. The global antigenic diversity of swine influenza A viruses. Elife 5:e12217. https://doi.org/10.7554/eLife.12217.
10. World Health Organization. 2019. Recommended composition of influenza virus vaccines for use in the 2019–2020 northern hemisphere influenza season. Wkly Epidemiol Rec 94:141–150. https://apps.who.int/iris/bitstream/handle/10665/311441/WER9412-141-150.pdf.
11. Zeller MA, Anderson TK, Walia RW, Vincent AL, Gauger PC. 2018. ISU FLUture: a veterinary diagnostic laboratory web-based platform to monitor the temporal genetic patterns of influenza A virus in swine. BMC Bioinformatics 19:397. https://doi.org/10.1186/s12859-018-2408-7.
12. Pardo FOC, Schelkopf A, Allerson M, Morrison R, Culhane M, Perez A, Torremorell M. 2018. Breed-to-wean farm factors associated with influenza A virus infection in piglets at weaning. Prev Vet Med 161:33–40. https://doi.org/10.1016/j.prevetmed.2018.10.008.
13. Bolton MJ, Abente EJ, Venkatesh D, Stratton JA, Zeller M, Anderson TK, Lewis NS, Vincent AL. 2019. Antigenic evolution of H3N2 influenza A viruses in swine in the United States from 2012 to 2016. Influenza Other Respir Viruses 13:83–90. https://doi.org/10.1111/irv.12610.
14. Abente EJ, Santos J, Lewis NS, Gauger PC, Stratton J, Skepner E, Anderson TK, Rajao DS, Perez DR, Vincent AL. 2016. The molecular determinants of antibody recognition and antigenic drift in the H3 hemagglutinin of swine influenza A virus. J Virol 90:8266–8280. https://doi.org/10.1128/JVI.01002-16.
15. Santos JJS, Abente EJ, Obadan AO, Thompson AJ, Ferreri L, Geiger G, Gonzalez-Reiche AS, Lewis NS, Burke DF, Rajão DS, Paulson JC, Vincent AL, Perez DR. 2018. Plasticity of amino acid residue 145 near the receptor binding site of H3 swine influenza A viruses and its impact on receptor binding and antibody recognition. J Virol 93:e01413-18. https://doi.org/10.1128/JVI.01413-18.
16. Das SR, Hensley SE, David A, Schmidt L, Gibbs JS, Puigbò P, Ince WL, Bennink JR, Yewdell JW. 2011. Fitness costs limit influenza A virus hemagglutinin glycosylation as an immune evasion strategy. Proc Natl Acad Sci U S A 108:E1417–E1422. https://doi.org/10.1073/pnas.1108754108.
17. Myers JL, Wetzel KS, Linderman SL, Li Y, Sullivan CB, Hensley SE. 2013. Compensatory hemagglutinin mutations alter antigenic properties of influenza viruses. J Virol 87:e01414-13. https://doi.org/10.1128/JVI.01414-13.
18. Li Y, Bostick DL, Sullivan CB, Myers JL, Griesemer SB, StGeorge K, Plotkin JB, Hensley SE. 2013. Single hemagglutinin mutations that alter both antigenicity and receptor binding avidity influence influenza virus antigenic clustering. J Virol 87:9904–9910. https://doi.org/10.1128/JVI.01023-13.
19. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. 2016. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci U S A 113:E1701–E1709. https://doi.org/10.1073/pnas.1525578113.
20. Eisler D, Fornika D, Tindale LC, Chan T, Sabaiduc S, Hickman R, Chambers C, Krajden M, Skowronski DM, Jassem A, Hsiao W. 2020. Influenza classification suite: an automated Galaxy workflow for rapid influenza sequence analysis. Influenza Other Respir Viruses 14:358–362. https://doi.org/10.1111/irv.12722.
21. Chang J, Anderson TK, Zeller MA, Gauger PC, Vincent AL. 2019. octoFLU: automated classification for the evolutionary origin of influenza A virus gene sequences detected in US swine. Microbiol Resour Announc 8:e00673-19. https://doi.org/10.1128/MRA.00673-19.
22. Breiman L, Friedman J, Stone CJ, Olshen RA. 1984. Classification and regression trees. CRC Press, Boca Raton, FL.
23. Yao Y, Li X, Liao B, Huang L, He P, Wang F, Yang J, Sun H, Zhao Y, Yang J. 2017. Predicting influenza antigenicity from hemagglutinin sequence data based on a joint random forest method. Sci Rep 7:1545. https://doi.org/10.1038/s41598-017-01699-z.
24. Sun H, Yang J, Zhang T, Long L-P, Jia K, Yang G, Webby RJ, Wan X-F. 2013. Using sequence data to infer the antigenicity of influenza virus. mBio 4:e00230-13. https://doi.org/10.1128/mBio.00230-13.
25. Yang J, Zhang T, Wan X-F. 2014. Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. PLoS One 9:e106660. https://doi.org/10.1371/journal.pone.0106660.
26. Harvey WT, Benton DJ, Gregory V, Hall JP, Daniels RS, Bedford T, Haydon DT, Hay AJ, McCauley JW, Reeve R. 2016. Identification of low-and high-impact hemagglutinin amino acid substitutions that drive antigenic drift of influenza A (H1N1) viruses. PLoS Pathog 12:e1005526. https://doi.org/10.1371/journal.ppat.1005526.
27. Ito K, Igarashi M, Miyazaki Y, Murakami T, Iida S, Kida H, Takada A. 2011. Gnarled-trunk evolutionary model of influenza A virus hemagglutinin. PLoS One 6:e25953. https://doi.org/10.1371/journal.pone.0025953.
28. Fitch WM, Bush RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci U S A 94:7712–7718. https://doi.org/10.1073/pnas.94.15.7712.
29. Nelson MI, Holmes EC. 2007. The evolution of epidemic influenza. Nat Rev Genet 8:196–205. https://doi.org/10.1038/nrg2053.
30. Anderson TK, Campbell BA, Nelson MI, Lewis NS, Janas-Martindale A, Killian ML, Vincent AL. 2015. Characterization of co-circulating swine influenza A viruses in North America and the identification of a novel H1

genetic clade with antigenic significance. Virus Res 201:24–31. https://doi.org/10.1016/j.virusres.2015.02.009.

31. Bell SM, Katzelnick L, Bedford T. 2019. Dengue genetic divergence generates within-serotype antigenic variation, but serotypes dominate evolutionary dynamics. Elife 8:e42496. https://doi.org/10.7554/eLife.42496.
32. Wiley D, Wilson I, Skehel J. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 289:373–378. https://doi.org/10.1038/289373a0.
33. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. 1999. Predicting the evolution of human influenza A. Science 286:1921–1925. https://doi.org/10.1126/science.286.5446.1921.
34. Fonville JM, Fraaij PL, de Mutsert G, Wilks SH, van Beek R, Fouchier RA, Rimmelzwaan GF. 2016. Antigenic maps of influenza A (H3N2) produced with human antisera obtained after primary infection. J Infect Dis 213:31–38. https://doi.org/10.1093/infdis/jiv367.
35. Breiman L. 2001. Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324.
36. Gao S, Anderson TK, Walia RR, Dorman KS, Janas-Martindale A, Vincent AL. 2017. The genomic evolution of H1 influenza A viruses from swine detected in the United States between 2009 and 2016. J Gen Virol 98:2001–2010. https://doi.org/10.1099/jgv.0.000885.
37. Torremorell M, Allerson M, Corzo C, Diaz A, Gramer M. 2012. Transmission of influenza A virus in pigs. Transbound Emerg Dis 59:68–84. https://doi.org/10.1111/j.1865-1682.2011.01300.x.
38. Zeller MA, Chang J, Vincent AL, Gauger PC, Anderson TK. 2020. Coordinated evolution between N2 neuraminidase and H1 and H3 hemagglutinin genes increased influenza A virus genetic diversity in swine. bioRxiv 2020.05.29.123828.
39. Ampofo WK, Baylor N, Cobey S, Cox NJ, Daves S, Edwards S, Ferguson N, Grohmann G, Hay A, Katz J, Kullabutr K, Lambert L, Levandowski R, Mishra AC, Monto A, Siqueira M, Tashiro M, Waddell AL, Wairagkar N, Wood J, Zambon M, Zhang W, WHO Writing Group. 2012. Improving influenza vaccine virus selection: report of a WHO informal consultation held at WHO headquarters, Geneva, Switzerland, 14–16 June 2010. Influenza Other Respir Viruses 6:142–152. https://doi.org/10.1111/j.1750-2659.2011.00277.x.
40. Pedersen JC. 2014. Hemagglutination-inhibition assay for influenza virus subtype identification and the detection and quantitation of serum antibodies to influenza virus. In Spackman E (ed), Animal influenza virus. Humana Press, New York, NY.
41. Kitikoon P, Gauger PC, Vincent AL. 2014. Hemagglutinin inhibition assay with swine sera, p 295–301. In Spackman E (ed), Animal influenza virus. Humana Press, New York, NY.
42. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, Fouchier RA. 2004. Mapping the antigenic and genetic evolution of influenza virus. Science 305:371–376. https://doi.org/10.1126/science.1097211.
43. Lewis NS, Daly JM, Russell CA, Horton DL, Skepner E, Bryant NA, Burke DF, Rash AS, Wood JLN, Chambers TM, Fouchier RAM, Mumford JA, Elton DM, Smith DJ. 2011. Antigenic and genetic evolution of equine influenza A (H3N8) virus from 1968 to 2007. J Virol 85:12742–12749. https://doi.org/10.1128/JVI.05319-11.
44. de Jong JC, Smith DJ, Lapedes AS, Donatelli I, Campitelli L, Barigazzi G, Van Reeth K, Jones TC, Rimmelzwaan GF, Osterhaus ADME, Fouchier RAM. 2007. Antigenic and genetic evolution of swine influenza A (H3N2) viruses in Europe. J Virol 81:4315–4322. https://doi.org/10.1128/JVI.02458-06.
45. Lewis NS, Anderson TK, Kitikoon P, Skepner E, Burke DF, Vincent AL. 2014. Substitutions near the hemagglutinin receptor-binding site determine the antigenic evolution of influenza A H3N2 viruses in U.S. swine. J Virol 88:4752–4763. https://doi.org/10.1128/JVI.03805-13.
46. Rajao DS, Gauger PC, Anderson TK, Lewis NS, Abente EJ, Killian ML, Perez DR, Sutton TC, Zhang J, Vincent AL. 2015. Novel reassortant human-like H3N2 and H3N1 influenza A viruses detected in pigs are virulent and antigenically distinct from swine viruses endemic to the United States. J Virol 89:11213–11222. https://doi.org/10.1128/JVI.01675-15.
47. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
48. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A. 2014. Integrating influenza antigenic dynamics with molecular evolution. Elife 3:e01914. https://doi.org/10.7554/eLife.01914.
49. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830.
50. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.
51. Rambaut A. 2012. FigTree v1. 4. Molecular evolution, phylogenetics and epidemiology. University of Edinburgh, Institute of Evolutionary Biology, Edinburgh, United Kingdom.
52. Lee PS, Ohshima N, Stanfield RL, Yu W, Iba Y, Okuno Y, Kurosawa Y, Wilson IA. 2014. Receptor mimicry by antibody F045–092 facilitates universal binding to the H3 subtype of influenza virus. Nat Commun 5:3614. https://doi.org/10.1038/ncomms4614.