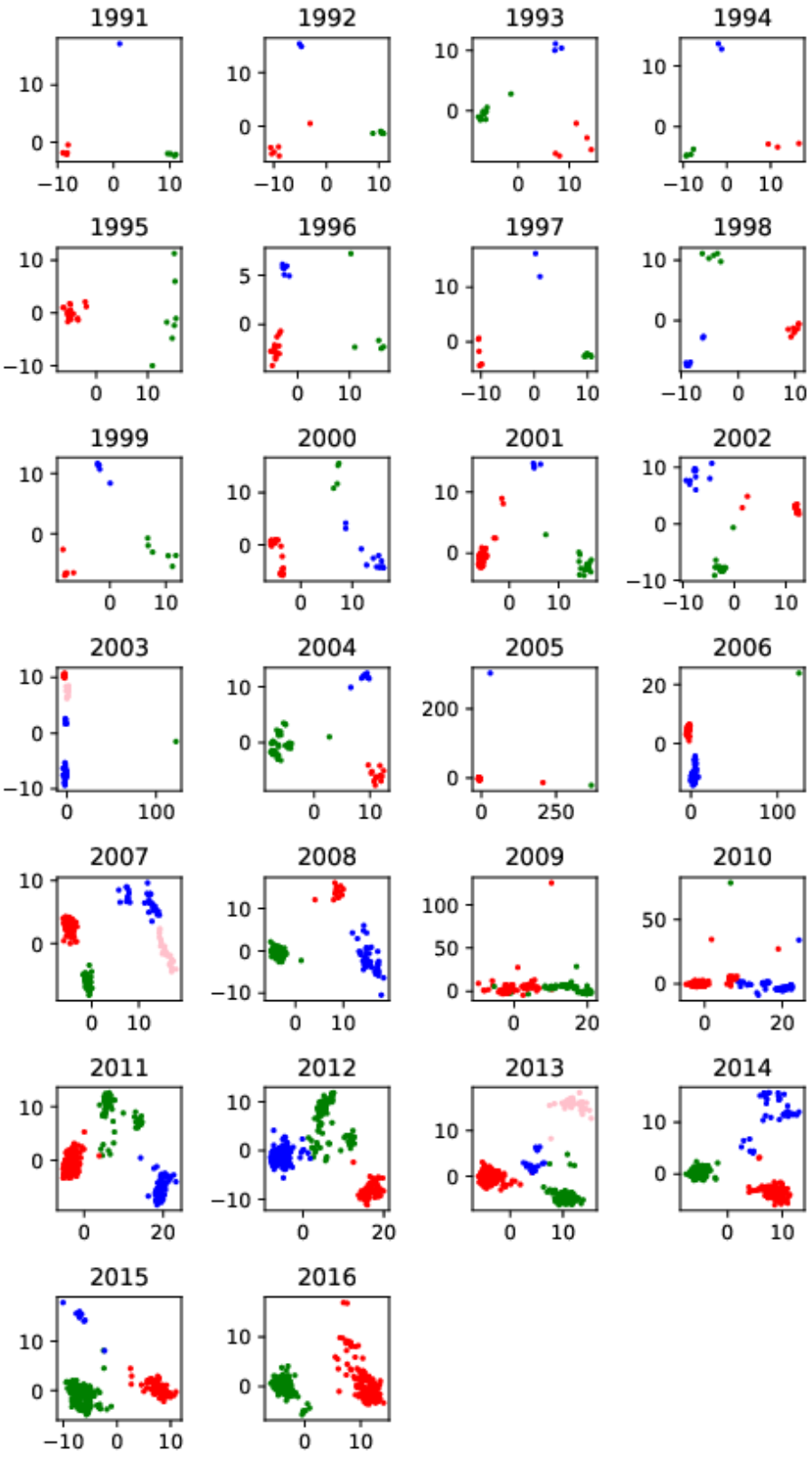
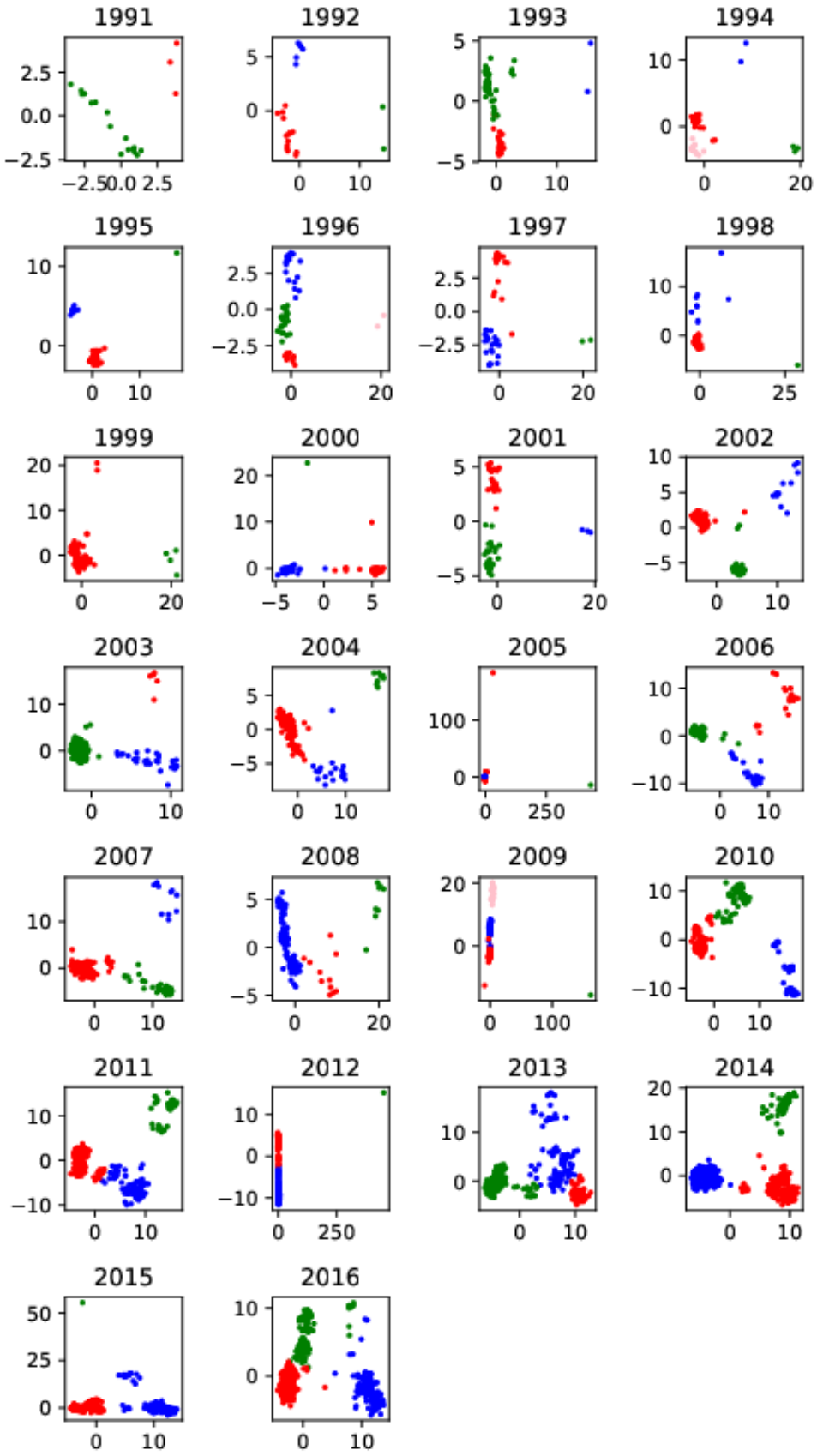
S1. The visualization of clusters of influenza strains on three subtypes H1N1, H3N2 and H5N1 across selected years.

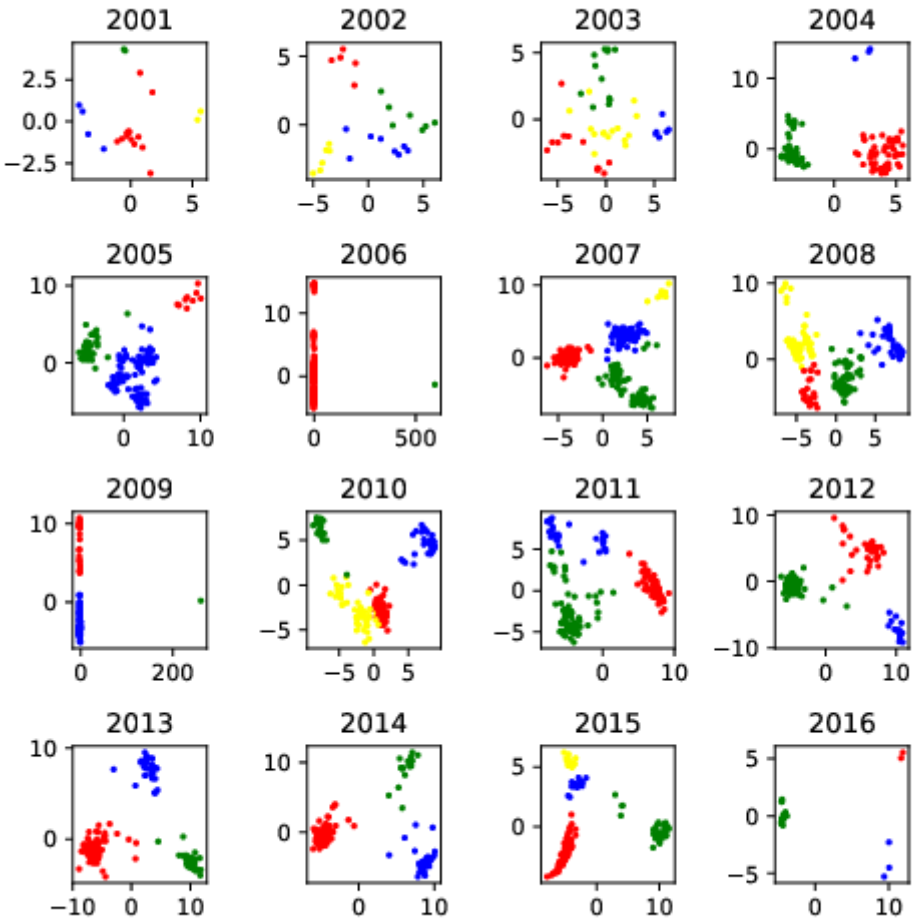
a) H1N1



b) H3N2


1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016

c) H5N1



The process of clustering of sequences from each year is described below:

1. The raw sequences are split into subsequences that each sequence consists of (n-2) lists of 3-grams as shown in Fig. 3, where n is the length of HA.
2. The subsequence of 3-grams is embedded into a 100-dimension vector based on ProtVec and each sequence is represented as the summation of the vector representation of subsequences. Thus, each sequence is presented as a vector of size 100.
3. The original raw sequence data is transformed into numerical data for all subtypes using ProtVec.
4. K-means is applied to the sequences from each year after embeddings and we can obtain the clusters of the sequences of each year.
5. We determine the choice of k in k-means with Elbow method by calculating the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the best k (1,2,3,4,5) for which WSS becomes first starts to diminish.
6. PCA is applied to the transformed numerical data that represents the sequences from different years and the visualization of clustering is presented for each year.

S2. The epitope sites for influenza subtypes H1N1, H3N2 and H5N1.

a) H1N1

Epitope	Sites
A	118, 120, 121, 122, 126, 127, 128, 129, 132, 133, 134, 135, 137, 139, 140, 141, 142, 143, 146, 147, 149, 165, 252, 253
B	124, 125, 152, 153, 154, 155, 156, 157, 160, 162, 163, 183, 184, 185, 186, 187, 189, 190, 191, 193, 194, 196
C	34, 35, 36, 37, 38, 40, 41, 43, 44, 45, 269, 270, 271, 272, 273, 274, 276, 277, 278, 283, 288, 292, 295, 297, 298, 302, 303, 305, 306, 307, 308, 309, 310
D	89, 94, 95, 96, 113, 117, 163, 164, 166, 167, 168, 169, 170, 171, 172, 173, 174, 176, 179, 198, 200, 202, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 222, 223, 224, 225, 226, 227, 235, 237, 239, 241, 243, 244, 245
E	47, 48, 50, 51, 53, 54, 56, 57, 58, 66, 68, 69, 70, 71, 72, 73, 74, 75, 78, 79, 80, 82, 83, 84, 85, 86, 102, 257, 258, 259, 260, 261, 263, 267

[1] Deem M W, Pan K. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy[J]. Protein Engineering Design and Selection, 2009, 22(9): 543-546.

b) H3N2

Epitope	Sites
A	122, 124, 126, 130, 131, 132, 133, 135, 136, 137, 138, 140, 142, 143, 144, 145, 146, 150, 152, 168
B	128, 129, 155, 156, 157, 158, 159, 160, 163, 165, 186, 187, 188, 189, 190, 192, 193, 194, 196, 197, 198
C	44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 273, 275, 276, 278, 279, 280, 294, 297, 299, 300, 304, 305, 307, 308, 309, 310, 311, 312
D	96, 102, 103, 117, 121, 167, 170, 171, 172, 173, 174, 175, 176, 177, 179, 182, 201, 203, 207, 208, 209, 212, 213, 214, 215, 216, 217, 218, 219, 226, 227, 228, 229, 230, 235, 238, 240, 242, 244, 246, 247, 248
E	57, 59, 62, 63, 67, 75, 78, 80, 81, 82, 83, 86, 87, 88, 91, 92, 94, 109, 260, 261, 262, 265

[2] Munoz E T, Deem M W. Epitope analysis for influenza vaccine design[J]. Vaccine, 2005, 23(9): 1144-1148.

c) H5N1

Sites on epitope	36, 48, 53, 55, 56, 57, 62, 65, 71, 77, 78, 80, 81, 82, 83, 84, 86, 87, 91, 94, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 133, 136, 138, 140, 141, 142, 143, 144, 145, 149, 150, 151, 152, 153, 154, 155, 156, 157, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 171, 172, 173, 174, 179, 182, 185, 186, 187, 189, 190, 191, 193, 200, 205, 206, 207, 212, 222, 226, 230, 242, 244, 245, 246, 252, 256, 259, 261, 262, 263, 273, 274, 276, 278, 282

[3] Li J, Wang Y, Liang Y, et al. Fine antigenic variation within H5N1 influenza virus hemagglutinin's antigenic sites defined by yeast cell surface display[J]. European journal of immunology, 2009, 39(12): 3498-3510.

[4] Kaverin N V, Rudneva I A, Govorkova E A, et al. Epitope mapping of the hemagglutinin molecule of a highly pathogenic H5N1 influenza virus by using monoclonal antibodies[J]. Journal of virology, 2007, 81(23): 12911-12917.

S3. The standard deviation of predictive results on test data of three influenza datasets at epitope sites.

Dataset	Model	Accuracy			Precision			Sensitivity			MCC		
		5	10	15	5	10	15	5	10	15	5	10	15
H1N1	Baseline	0.004	0.003	0.001	0.019	0.012	0.015	0.015	0.013	0.007	0.009	0.012	0.016
	LR	0.002	0.004	0.001	0.004	0.005	0.06	0.005	0.005	0.004	0.002	0.005	0.004
	SVM	0.001	0.002	0.001	0.005	0.001	0.002	0.006	0.006	0.006	0.001	0.002	0.002
	RNN	0.002	0.002	0.001	0.006	0.008	0.005	0.007	0.005	0.005	0.002	0.002	0.002
	GRU	0.002	0.002	0.001	0.010	0.006	0.008	0.009	0.006	0.008	0.002	0.001	0.003
	LSTM	0.002	0.002	0.001	0.008	0.006	0.007	0.009	0.006	0.008	0.002	0.002	0.002
	Tempel	0.003	0.001	0.001	0.007	0.005	0.006	0.008	0.007	0.008	0.002	0.001	0.002
H3N2	Baseline	0.002	0.005	0.003	0.005	0.011	0.007	0.003	0.002	0.005	0.010	0.012	0.009
	LR	0.002	0.003	0.002	0.008	0.009	0.008	0.005	0.008	0.006	0.002	0.003	0.005
	SVM	0.003	0.001	0.002	0.009	0.007	0.008	0.005	0.010	0.007	0.005	0.004	0.008
	RNN	0.003	0.002	0.002	0.011	0.008	0.009	0.005	0.006	0.006	0.007	0.010	0.004
	GRU	0.002	0.003	0.002	0.005	0.005	0.004	0.003	0.007	0.006	0.006	0.009	0.005
	LSTM	0.002	0.003	0.002	0.005	0.004	0.004	0.004	0.005	0.004	0.005	0.008	0.005
	Tempel	0.002	0.002	0.002	0.004	0.003	0.003	0.004	0.005	0.003	0.005	0.008	0.006
H5N1	Baseline	0.010	0.001	0.002	0.011	0.003	0.010	0.005	0.004	0.005	0.002	0.005	0.003
	LR	0.002	0.001	0.001	0.002	0.002	0.001	0.001	0.001	0.001	0.002	0.002	0.002
	SVM	0.001	0.002	0.002	0.002	0.001	0.003	0.000	0.000	0.000	0.001	0.001	0.002
	RNN	0.001	0.001	0.003	0.003	0.002	0.004	0.002	0.003	0.003	0.003	0.004	0.003
	GRU	0.001	0.002	0.001	0.002	0.003	0.002	0.003	0.003	0.002	0.002	0.003	0.002
	LSTM	0.001	0.001	0.001	0.002	0.002	0.002	0.003	0.001	0.003	0.002	0.001	0.003
	Tempel	0.001	0.001	0.002	0.002	0.003	0.002	0.002	0.001	0.001	0.003	0.003	0.002

S4. The prediction results on three influenza datasets at epitope sites using the training set.

Dataset	Model	Accuracy			Precision			Sensitivity			MCC		
		5	10	15	5	10	15	5	10	15	5	10	15
H1N1	Baseline	0.850	0.859	0.876	0.495	0.538	0.617	0.360	0.352	0.412	0.343	0.364	0.449
	LR	0.884	0.884	0.885	0.687	0.687	0.688	0.377	0.383	0.393	0.466	0.467	0.467
	SVM	0.881	0.878	0.879	0.570	0.562	0.575	0.814	0.811	0.820	0.615	0.610	0.612
	RNN	0.946	0.947	0.946	0.846	0.848	0.847	0.778	0.805	0.806	0.779	0.784	0.782
	GRU	0.950	0.950	0.951	0.856	0.858	0.859	0.783	0.814	0.814	0.799	0.801	0.802
	LSTM	0.952	0.952	0.951	0.856	0.850	0.855	0.789	0.820	0.823	0.800	0.801	0.805
	Tempel	0.952	0.954	0.953	0.857	0.860	0.860	0.791	0.821	0.825	0.802	0.805	0.806
H3N2	Baseline	0.936	0.913	0.897	0.464	0.333	0.285	0.430	0.452	0.457	0.405	0.344	0.316
	LR	0.940	0.940	0.941	0.466	0.462	0.464	0.045	0.043	0.044	0.129	0.130	0.128
	SVM	0.778	0.781	0.783	0.187	0.189	0.191	0.814	0.814	0.815	0.318	0.321	0.322
	RNN	0.959	0.956	0.953	0.762	0.730	0.791	0.440	0.457	0.455	0.610	0.565	0.555
	GRU	0.960	0.961	0.960	0.764	0.779	0.768	0.540	0.567	0.539	0.610	0.629	0.624
	LSTM	0.961	0.961	0.962	0.777	0.774	0.777	0.550	0.573	0.634	0.617	0.627	0.629
	Tempel	0.961	0.964	0.963	0.776	0.784	0.782	0.551	0.578	0.644	0.625	0.635	0.633
H5N1	Baseline	0.960	0.973	0.959	0.163	0.233	0.161	0.332	0.356	0.383	0.240	0.284	0.238
	LR	0.987	0.987	0.987	0.840	0.847	0.862	0.157	0.145	0.145	0.380	0.382	0.408
	SVM	0.927	0.924	0.930	0.166	0.161	0.166	0.949	0.949	0.949	0.382	0.374	0.385
	RNN	0.987	0.987	0.983	0.847	0.817	0.810	0.442	0.439	0.425	0.628	0.630	0.621
	GRU	0.988	0.988	0.989	0.876	0.849	0.844	0.486	0.487	0.497	0.663	0.659	0.661
	LSTM	0.988	0.988	0.988	0.859	0.844	0.854	0.489	0.493	0.484	0.649	0.651	0.646
	Tempel	0.991	0.992	0.991	0.885	0.890	0.886	0.475	0.485	0.489	0.669	0.678	0.670

S5. The results of mutation prediction at the single epitope site on three influenza datasets using our proposed framework.

Dataset*	Sites**	Accuracy	Precision	F-measure	MCC
	35	0.951	0.913	0.655	0.662
	51	0.883	0.885	0.890	0.766
	70	0.922	0.951	0.713	0.701
	72	0.970	0.981	0.873	0.863
	73	0.904	0.860	0.917	0.814
	82	1.000	1.000	1.000	1.000
	84	0.734	0.627	0.733	0.511
	85	0.974	0.978	0.979	0.943
	89	0.926	0.973	0.593	0.617
	96	0.949	0.923	0.949	0.899
	102	0.810	0.790	0.857	0.601
	113	0.916	0.888	0.930	0.831
	118	1.000	1.000	1.000	1.000
	128	1.000	1.000	0.995	0.995
	129	0.809	0.732	0.802	0.631
	137	0.913	0.909	0.889	0.823

H1N1	140	0.994	1.000	0.764	0.783
	141	0.986	1.000	0.567	0.625
	143	0.789	0.718	0.801	0.601
	146	0.719	0.643	0.739	0.471
	149	0.974	0.980	0.789	0.793
	153	0.848	0.787	0.845	0.704
	154	0.916	0.883	0.918	0.836
	155	0.943	0.932	0.953	0.882
	157	0.884	0.850	0.901	0.770
	165	0.895	0.869	0.913	0.789
	167	0.948	0.924	0.953	0.896
	168	0.885	0.866	0.899	0.769
	169	0.887	0.689	0.564	0.512
	171	0.811	0.917	0.567	0.566
	172	0.956	0.971	0.958	0.911
	176	0.974	0.980	0.789	0.793
	184	0.948	0.960	0.649	0.665
	186	0.712	0.610	0.711	0.564
	187	0.980	0.923	0.469	0.456
	194	0.828	0.749	0.843	0.684
	200	0.942	0.950	0.619	0.638
	202	0.932	0.924	0.948	0.853
	206	0.827	0.866	0.717	0.616
	207	0.999	1.000	0.933	0.935
	209	0.910	0.676	0.577	0.535
	210	0.975	0.990	0.797	0.801
	211	0.954	0.969	0.961	0.906
	212	0.716	0.791	0.494	0.433
	213	0.815	0.949	0.514	0.543
	215	0.984	1.000	0.441	0.527
	216	0.890	0.839	0.904	0.790
	223	0.941	0.920	0.947	0.883
	224	0.825	0.815	0.435	0.456
	225	0.828	0.749	0.832	0.677
	227	0.896	0.867	0.912	0.792
	235	1.000	1.000	1.000	1.000
	243	0.973	0.970	0.876	0.866
	252	0.948	0.924	0.952	0.896
	257	0.923	0.921	0.547	0.570
	261	0.942	0.940	0.616	0.633
	269	0.969	0.969	0.750	0.756
	272	0.983	0.994	0.901	0.896

	274	0.960	0.969	0.965	0.917
	276	0.828	0.751	0.845	0.685
	277	0.936	0.966	0.644	0.858
	278	0.969	0.968	0.973	0.937
	283	0.917	0.893	0.930	0.833
	288	0.934	0.936	0.609	0.624
	292	0.884	0.875	0.698	0.651
	302	0.968	0.955	0.966	0.935
	303	0.967	0.941	0.739	0.742
	305	0.969	0.949	0.855	0.842
	46	0.940	0.932	0.506	0.574
	47	0.990	1.000	0.583	0.683
	48	0.936	0.807	0.592	0.585
	49	0.972	0.875	0.300	0.409
	63	0.911	0.891	0.647	0.631
	78	0.980	0.935	0.683	0.701
	82	0.991	1.000	0.642	0.684
	94	0.984	1.000	0.576	0.654
	96	0.996	0.943	0.589	0.647
	102	0.990	1.000	0.512	0.507
	103	0.978	1.000	0.526	0.591
	109	0.951	0.875	0.718	0.705
	121	0.971	1.000	0.554	0.610
	126	0.981	0.921	0.648	0.671
	129	0.996	0.994	0.972	0.970
	132	0.871	0.506	0.410	0.448
	136	0.673	0.753	0.367	0.335
	137	0.955	0.977	0.790	0.784
	146	0.876	0.847	0.430	0.479
	150	0.912	0.907	0.627	0.621
	152	0.908	0.891	0.651	0.633
	155	0.814	0.749	0.445	0.402
	157	0.663	0.737	0.532	0.430
	158	0.973	0.830	0.615	0.625
	159	0.809	0.812	0.638	0.538
	160	0.813	0.586	0.624	0.502
	165	0.985	0.983	0.786	0.796
	170	0.946	0.798	0.647	0.632
	171	0.899	0.686	0.789	0.739
	172	0.975	0.952	0.860	0.851
	173	0.920	0.930	0.684	0.673
	174	0.879	0.815	0.754	0.678

H3N2	175	0.903	0.910	0.716	0.683
	179	0.874	0.506	0.496	0.440
	182	0.989	1.000	0.531	0.597
	186	0.645	0.510	0.641	0.480
	187	0.987	0.942	0.784	0.789
	188	0.947	0.981	0.747	0.746
	190	0.994	1.000	0.629	0.675
	197	0.986	1.000	0.540	0.603
	201	0.944	0.864	0.659	0.652
	203	0.942	0.898	0.475	0.519
	207	0.980	0.945	0.857	0.851
	208	0.824	0.664	0.693	0.570
	212	0.876	0.944	0.686	0.654
	213	0.856	0.862	0.475	0.475
	214	0.905	0.787	0.555	0.537
	216	0.973	0.938	0.738	0.743
	217	0.973	0.917	0.549	0.514
	218	0.844	0.569	0.547	0.453
	219	0.946	0.864	0.511	0.565
	227	0.958	0.769	0.541	0.547
	228	0.976	0.945	0.684	0.702
	229	0.970	1.000	0.719	0.737
	230	0.987	0.933	0.509	0.567
	235	1.000	1.000	1.000	1.000
	238	0.941	0.813	0.694	0.671
	240	0.982	0.945	0.852	0.897
	242	0.990	0.814	0.778	0.773
	244	0.985	1.000	0.544	0.530
	247	0.993	0.923	0.828	0.828
	248	0.976	0.941	0.662	0.683
	261	0.984	1.000	0.667	0.701
	275	0.956	0.954	0.582	0.616
	276	0.795	0.903	0.561	0.512
	278	0.989	1.000	0.703	0.732
	279	0.994	1.000	0.875	0.879
	294	0.994	1.000	0.698	0.730
	297	0.988	1.000	0.556	0.616
	304	0.902	0.667	0.457	0.435
	305	0.995	1.000	0.783	0.800
	307	0.980	0.857	0.574	0.523
	36	1.000	1.000	1.000	1.000
	62	1.000	1.000	1.000	1.000

H5N1	81	0.959	0.676	0.539	0.531
	83	0.965	0.684	0.594	0.582
	86	0.959	0.573	0.534	0.514
	122	1.000	1.000	1.000	1.000
	130	0.963	0.765	0.711	0.693
	138	0.920	1.000	0.024	0.107
	142	0.926	0.928	0.747	0.725
	144	1.000	1.000	1.000	1.000
	149	1.000	1.000	1.000	1.000
	151	0.889	0.850	0.594	0.600
	156	1.000	1.000	1.000	1.000
	168	1.000	1.000	1.000	1.000
	169	0.962	0.669	0.724	0.706
	171	0.959	0.730	0.689	0.668

*For every selected single site, we generate 10000 samples of 3-3-grams to predict mutations in the year 2016 with Tempel. The model is trained using sequential data from 2006-2015. The constructed samples are divided into training and testing set in a ratio of 0.8:0.2.

** This table shows the predictive performance on the testing set. Only the sites that the mutated samples exceed 1% of the testing set are presented.