Supplementary Information for Manuscript Entitled:
Accurate Virus Identification with Interpretable Raman Signatures by Machine Learning

Jiarong Ye[1], Yin-Ting Yeh[2], Yuan Xue[3], Ziyang Wang[4], Na Zhang[2], He Liu[2], Kunyan Zhang[4], RyeAnne Ricker[5], Zhuohang Yu[2], Allison Roder[6], Nestor Perea Lopez[2], Lindsey Organtini[7], Wallace Greene[8], Susan Hafenstein[7], Huaguang Lu[9], Elodie Ghedin[6], Mauricio Terrones[2], Shengxi Huang[4], Sharon Xiaolei Huang[1,*]

[1.] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA, USA
[2.] Department of Physics, The Pennsylvania State University, University Park, PA, USA
[3.] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
[4.] Department of Electrical and Computer Engineering, The Pennsylvania State University, University Park, PA, USA
[5] Department of Biomedical Engineering, George Washington University, Washington, D.C., USA
[6.] Systems Genomics Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA
[7.] Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA, USA
[8.] Department of Pathology and Laboratory Medicine, Division of Clinical Pathology, The Pennsylvania State University College of Medicine, Hershey, PA, USA
[9.] Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, PA, USA

Corresponding Author(s):

Sharon Xiaolei Huang
The Pennsylvania State University
University Park, PA 16802, USA
Email: suh972@psu.edu

This PDF file includes:

Figures S1 to S10
Tables S1 to S4

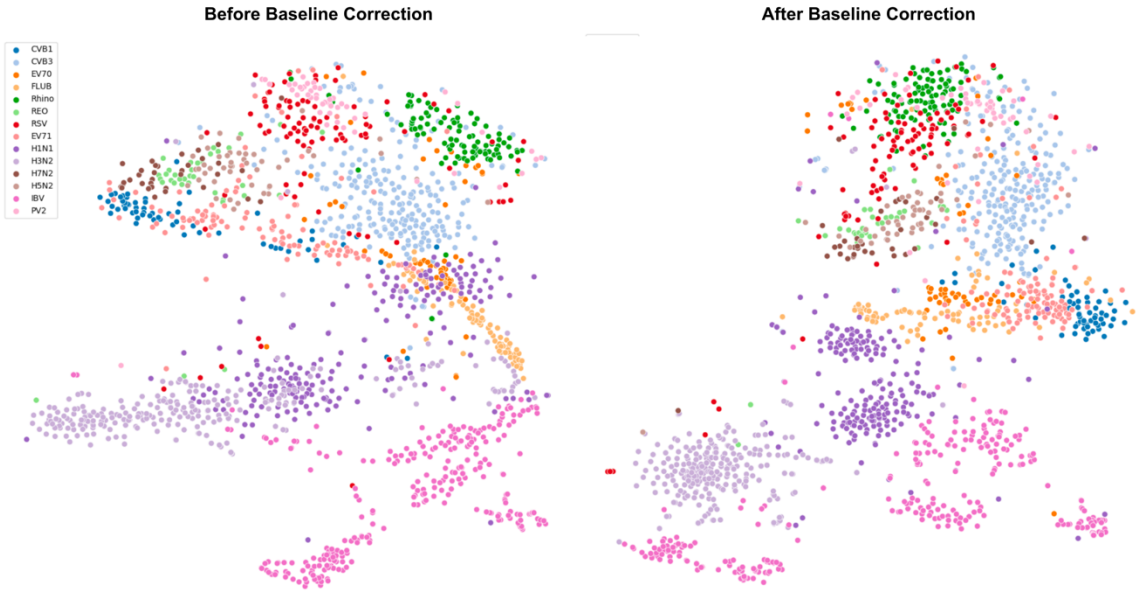

Fig. S1. The t-SNE plots of all viruses (Avian virus, Enterovirus and Human Respiratory viruses), before and after baseline correction. Each Raman spectrum is represented by a point in the plots. Observed from the comparison between the two plots, applying baseline correction makes the spectra of virus types (or subtypes) such as H3N2, H7N2, CVB1, RSV, EV71 more distinguishable by pulling tighter each cluster corresponding to spectra of the same virus while pushing the clusters of different viruses further apart.

A

Metrics	Avian Virus
Accuracy	0.9632 ± 0.0183
Sensitivity	0.8916 ± 0.0570
Specificity	0.9768 ± 0.0154

B

	IBV Coronavirus	Avian Influenza A Virus	Reovirus
Accuracy	0.9981 ± 0.0050	0.9088 ± 0.0757	0.7679 ± 0.1757

C

		IBV Coronavirus	Avian Influenza A Virus	Reovirus
Lipid	Phosphatidylcholine, Phosphatidylethanolamine, Sphingomyelin	58.19%	50.85%	52.54%
Protein	Amide I	64.29%	42.86%	67.86%
	Amide III	39.47%	26.32%	23.68%
Nucleic Acid	RNA	36.25%	48.75%	56.25%
Amino Acid	Tyrosine	60.00%	65.00%	48.00%
	Phenylalanine	52.50%	51.25%	42.50%
Other Functional Groups	C-C aliphatic chains	70.16%	83.06%	74.35%
	C-CH3	0.00%	30.00%	26.67%
	CH2	94.00%	56.00%	48.00%
	CH3	94.00%	56.00%	48.00%
	Carboxylate salt	67.50%	49.17%	33.33%
	Carboxylic acid	45.38%	33.08%	48.46%
	Ketone	31.82%	20.91%	30.00%

Fig. S2. A. The CNN classification performance of Avian viruses on three metrics (Accuracy, Sensitivity and Specificity); **B.** The CNN classification accuracy for each type of Avian virus; **C.** Matching scores between Raman ranges important for identifying Avian viruses using ML and Raman peak ranges of biomolecules.

A

Metrics	Enterovirus
Accuracy	0.9417 ± 0.0233
Sensitivity	0.9312 ± 0.0335
Specificity	0.9850 ± 0.0065

B

	CVB1	CVB3	EV70	EV71	PV2
Accuracy	0.9089 ± 0.1127	0.9768 ± 0.0200	0.9117± 0.0857	0.8805 ± 0.0637	0.9780 ± 0.0460

C

		CVB1	CVB3	EV70	EV71	PV2
Lipid	Phosphatidylcholine, Phosphatidylethanolamine, Sphingomyelin	58.19%	50.85%	52.54%	57.63%	57.63%
Protein	Amide I	64.29%	42.86%	67.86%	89.29%	7.14%
	Amide III	39.47%	26.32%	23.68%	13.16%	39.47%
Nucleic Acid	RNA	36.25%	48.75%	56.25%	68.75%	36.25%
Amino Acid	Tyrosine	60.00%	65.00%	48.00%	52.00%	56.00%
	Phenylalanine	52.50%	51.25%	42.50%	56.25%	56.88%
Other Functional Groups	C-C aliphatic chains	70.16%	83.06%	74.35%	74.68%	86.29%
	C-CH3	0.00%	30.00%	26.67%	0.00%	36.67%
	CH2	94.00%	56.00%	48.00%	66.00%	58.00%
	CH3	94.00%	56.00%	48.00%	66.00%	58.00%
	Carboxylate salt	67.50%	49.17%	33.33%	37.50%	31.67%
	Carboxylic acid	45.38%	33.08%	48.46%	55.38%	32.31%
	Ketone	31.82%	20.91%	30.00%	39.09%	11.82%

Fig. S3. A. The CNN classification performance of Enteroviruses on three metrics (Accuracy, Sensitivity and Specificity); **B.** The CNN classification accuracy of each type (subtype) of Enterovirus; **C.** Matching scores between Raman ranges important for identifying Enteroviruses using ML and Raman peak ranges of biomolecules.

A

Metrics	Influenza A Subtypes
Accuracy	0.9648 ± 0.0113
Sensitivity	0.8949 ± 0.0399
Specificity	0.9891 ± 0.0040

B

	H1N1	H3N2	H5N2	H7N2
Accuracy	0.9884 ± 0.0179	0.9922 ± 0.0111	0.8960 ± 0.1121	0.7029 ± 0.1452

C

		H1N1	H3N2	H5N2	H7N2
Lipid	Phosphatidylcholine, Phosphatidylethanolamine, Sphingomyelin	62.71%	70.62%	61.58%	50.28%
Protein	Amide I	75.00%	32.14%	57.14%	57.14%
	Amide III	10.53%	26.32%	0.00%	0.00%
Nucleic Acid	RNA	45.00%	51.25%	50.00%	58.75%
Amino Acid	Tyrosine	70.00%	89.00%	66.00%	73.00%
	Phenylalanine	56.88%	56.88%	67.50%	63.13%
Other Functional Groups	C-C aliphatic chains	70.97%	64.03%	79.03%	77.42%
	C-CH3	0.00%	30.00%	0.00%	0.00%
	CH2	78.00%	88.00%	100%	94.00%
	CH3	78.00%	88.00%	100%	94.00%
	Carboxylate salt	44.17%	68.33%	44.17%	50.83%
	Carboxylic acid	42.31%	46.15%	45.38%	36.92%
	Ketone	38.18%	33.64%	29.09%	30.91%

Fig. S4. A. The CNN classification performance of FLU A virus subtypes on three metrics (Accuracy, Sensitivity and Specificity); **B.** The CNN classification accuracy of each subtype of FLU A virus; **C.** Matching scores between Raman ranges important for identifying Influenza A subtypes using ML and Raman peak ranges of biomolecules.

A

	Avian fluA, Human fluA	Avian fluA, Human fluA, Human flu B	Human fluA, Human fluB
Accuracy	0.9961 ± 0.0056	0.9916 ± 0.0085	0.9947 ± 0.0071
Sensitivity	0.9907 ± 0.0177	0.9811 ± 0.0220	0.9813 ± 0.0250
Specificity	0.9907 ± 0.0177	0.9909 ± 0.0105	0.9813 ± 0.0250

B

		Avian fluA, Human fluA	Avian fluA, Human fluA, Human flu B	Human fluA, Human fluB
Lipid	Phosphatidylcholine, Phosphatidylethanolamine, Sphingomyelin	74.58%	54.80%	40.68%
Protein	Amide I	64.29%	32.14%	28.57%
	Amide III	13.16%	52.63%	73.68%
Nucleic Acid	RNA	40.00%	62.50%	60.00%
Amino Acid	Tyrosine	78.00%	69.00%	62.00%
	Phenylalanine	73.75%	63.13%	56.25%
Other Functional Groups	C-C aliphatic chains	62.10%	69.19%	73.06%
	C-CH3	0.00%	13.33%	20.00%
	CH2	66.00%	98.00%	60.00%
	CH3	66.00%	98.00%	60.00%
	Carboxylate salt	34.17%	44.17%	41.67%
	Carboxylic acid	73.85%	45.38%	32.31%
	Ketone	71.82%	48.18%	38.18%

Fig. S5. A. The CNN performance on three classification tasks involving Avian and Human flu viruses (1. Avian FLUA vs. Human FLUA; 2. Avian FLUA, Human FLUA, Human FLUB; 3. Human FLUA vs. Human FLUB); **B.** Matching scores between Raman ranges important for each of the three classification tasks using ML and Raman peak ranges of biomolecules.

A

Metrics	Within Enveloped	Within Non-Enveloped	Between Enveloped and Non-Enveloped
Accuracy	0.9751± 0.0090	0.9539 ± 0.0145	0.9477 ± 0.0081
Sensitivity	0.9711 ± 0.0146	0.9516 ± 0.0173	0.9474 ± 0.0079
Specificity	0.9923 ± 0.0029	0.9920 ± 0.0025	0.9474 ± 0.0079

B

		Within Enveloped	Within Non-Enveloped	Between Enveloped and Non-Enveloped
Lipid	Phosphatidylcholine, Phosphatidylethanolamine, Sphingomyelin	38.98%	51.98%	51.98%
Protein	Amide I	50.00%	39.29%	25.00%
	Amide III	5.26%	15.79%	7.89%
Nucleic Acid	RNA	50.00%	45.00%	41.25%
Amino Acid	Tyrosine	69.00%	58.00%	56.00%
	Phenylalanine	47.50%	60.00%	49.38%
Other Functional Groups	C-C aliphatic chains	55.97%	61.61%	82.10%
	C-CH3	3.33%	0.00%	0.00%
	CH2	82.00%	76.00%	0.00%
	CH3	82.00%	76.00%	0.00%
	Carboxylate salt	40.00%	41.67%	0.00%
	Carboxylic acid	60.77%	38.46%	37.69%
	Ketone	54.55%	25.45%	50.91%

Fig. S6. A. The CNN performance on three classification tasks involving enveloped and non-enveloped viruses (1. Classification within enveloped viruses, including FLUA, FLUB, IBV, RSV; 2. Classification within non-enveloped viruses, including Reovirus, Enterovirus, Rhino; 3. Binary classification to identify a virus as either enveloped or non-enveloped; **B.** Matching scores between Raman ranges important for each of the three classification tasks using ML and Raman peak ranges of biomolecules.

A

Metrics	Human Respiratory
Accuracy	0.9390 ± 0.0178
Sensitivity	0.8856 ± 0.0344
Specificity	0.9809 ± 0.0061

B

	Human fluA	Human fluB	Rhino	RSV
Accuracy	0.9938 ± 0.0066	0.7487 ± 0.0997	0.8250 ± 0.1016	0.9751 ± 0.035

C

		Human fluA	Human fluB	Rhino	RSV
Lipid	Phosphatidylcholine, Phosphatidylethanolamine, Sphingomyelin	71.19%	37.85%	41.81%	66.67%
Protein	Amide I	89.29%	0.00%	0.00%	92.86%
	Amide III	60.53%	60.53%	63.16%	73.68%
Nucleic Acid	RNA	32.50%	50.00%	60.00%	56.25%
Amino Acid	Tyrosine	57.00%	81.00%	76.00%	69.00%
	Phenylalanine	59.38%	77.50%	65.63%	50.00%
Other Functional Groups	C-C aliphatic chains	63.55%	71.61%	72.26%	66.29%
	C-CH3	0.00%	33.33%	40.00%	36.67%
	CH2	72.00%	42.00%	20.00%	90.00%
	CH3	72.00%	42.00%	20.00%	90.00%
	Carboxylate salt	19.17%	69.17%	65.83%	54.17%
	Carboxylic acid	81.54%	33.85%	31.54%	66.15%
	Ketone	78.18%	25.45%	22.73%	64.55%

Fig. S7. A. The CNN classification performance of Human Respiratory viruses on three metrics (Accuracy, Sensitivity and Specificity); **B.** The CNN classification accuracy for each type of Human Respiratory virus; **C.** Matching scores between Raman ranges important for identifying different types of Human Respiratory viruses using ML and Raman peak ranges of biomolecules.

A

	Accuracy	Sensitivity	Specificity
All	0.9224 ± 0.0114	0.8758 ± 0.0203	0.9929 ± 0.0010

B

	IBV Coronavirus	Avian fluA Virus	Reovirus	CVB1	CVB3	EV70	EV71	PV2	Human fluA	Human fluB	Rhino	RSV
Accuracy	0.9833 ± 0.0138	0.9476 ± 0.0490	0.7529 ± 0.1508	0.8856 ± 0.0808	0.9660 ± 0.0245	0.7900 ± 0.1057	0.7961 ± 0.0758	0.968 ± 0.0466	0.9784 ± 0.0153	0.8741 ± 0.0662	0.818 ± 0.0619	0.7506 ± 0.0957

Fig. S8. A. The overall CNN performance of classifying / identifying virus type (subtype) among all viruses in our dataset in one classification task; **B.** The classification accuracy for each type of virus, including Avian, Enterovirus and Human Respiratory viruses.

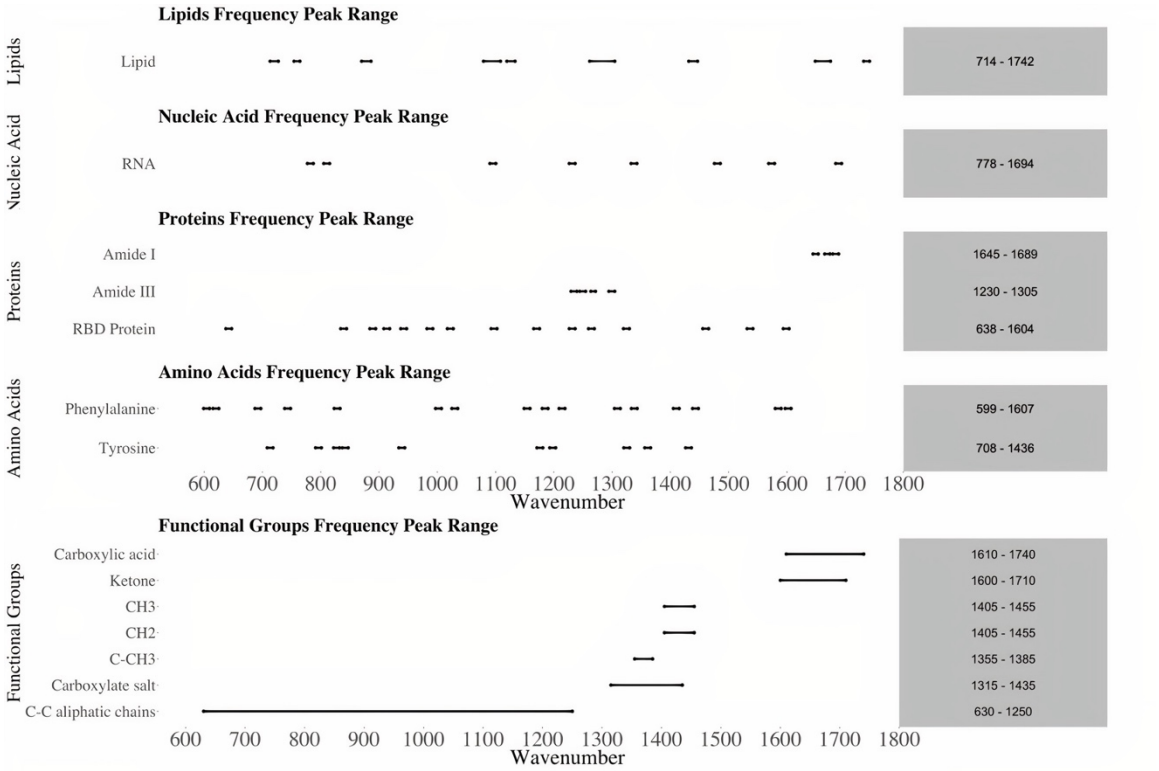

Fig. S9. Raman peak ranges of lipids (phosphatidylcholine, phosphatidylethanolamine and sphingomyelin), nucleic acids, proteins, amino acids and other chemical functional groups such as Carboxylic acid and Ketone. These peak ranges are used for matching score calculation to help us understand what biomolecules or chemical functional groups are important for virus identification tasks using ML.

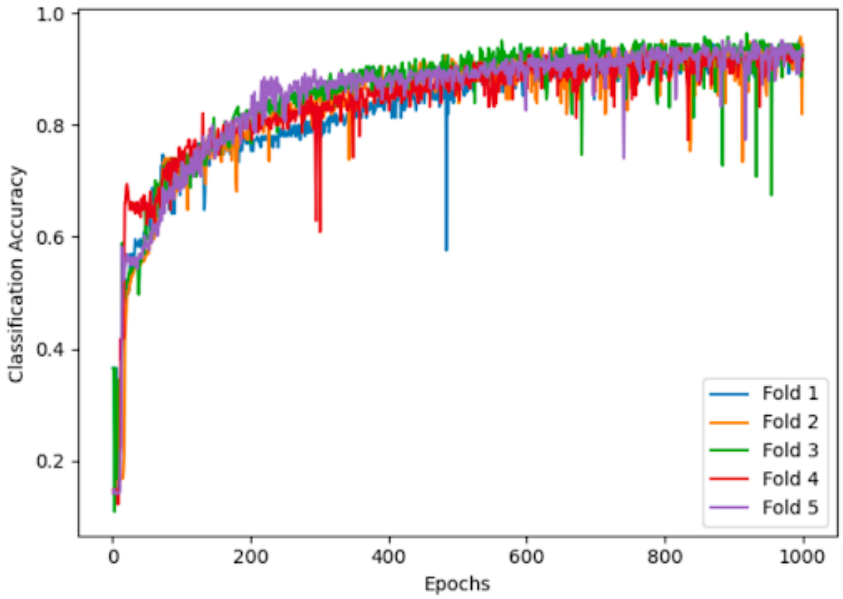

Fig. S10. Learning curves of 5-fold cross validation for the classification task on Flu A subtypes (H1N1/H3N2/H5N2/H7N2). Each of the five folds is used as the hold-out validation set once, and the learning curves for the validation folds are shown in the figure. In each learning curve, the classification accuracy on the validation fold after each training epoch is plotted. Although with some fluctuations, the learning curves for the five folds are similar and they all converge when the training process gets close to 1000 epochs, which justifies our choice for the number of training epochs, one among many crucial hyper-parameters.

Table S1: Definition for ML classification performance metrics: Sensitivity, Specificity and Accuracy. Sensitivity is the percentage of positive cases correctly identified as positive. Specificity is the percentage of negative cases correctly identified as negative. Accuracy is the percentage of correctly identified cases.

		Predicted Class		
Actual Class		**Positive**	**Negative**	
	Positive	True Positive (TP)	False Negative (FN)	$\text{Sensitivity} = \frac{TP}{TP+FN}$
	Negative	False Positive (FP)	True Negative (TN)	$\text{Specificity} = \frac{TN}{TN+FP}$
				$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$

Table S2. Information about a large dataset consisting of Raman spectra of various types of flu viruses, which is used to test the viral dose detection limit of our approach. For each flu virus strain, we have collected around 10,000 Raman spectra.

Sample ID	Flu Virus Strain	Flu Type/Subtype
1	A/North Carolina/04/2016	Flu A / H3N2
2	A/Nebraska/14/2019	Flu A / H1N1
3	B/Massachusetts/02/2012	Flu B
4	A/Michigan/45/2015	Flu A / H1N1
5	A/Hawaii/47/2014	Flu A / H3N2
6	A/California/07/2009	Flu A / H1N1
7	A/Indiana/08/2018	Flu A / H3N2
8	A/Arizona/45/2018	Flu A / H3N2
9	A/Delaware/39/2019	Flu A / H3N2
10	A/Singapore/INFIMH-16_0010/2016	Flu A / H3N2
11	A/Idaho/07/2018	Flu A / H1N1

Table S3: The TCID50 and RNA copies present in 10 µL of sample, the volume used for spectra collection.

	Flu A/Nebraska/14/2019 (H1N1)		Flu A/Indiana/08/2018 (H3N2)	
Dilution	**TCID50/10µL**	**RNA copies/10µL**	**TCID50/10µL**	**RNA copies/10µL**
Undiluted	2.29×10^{5}	2.27×10^{7}	1.45×10^{5}	1.42×10^{7}
10^{-1}	2.29×10^{4}	2.27×10^{6}	1.45×10^{4}	1.42×10^{6}
10^{-2}	2.29×10^{3}	2.27×10^{5}	1.45×10^{3}	1.42×10^{5}
10^{-3}	2.29×10^{2}	2.27×10^{4}	1.45×10^{2}	1.42×10^{4}
10^{-4}	2.29×10^{1}	2.27×10^{3}	1.45×10^{1}	1.42×10^{3}
10^{-5}	2	227	1	142
10^{-6}	<1	23	<1	14

Table S4: Accuracy of flu type and subtype classification for two testing strains, Indiana/08 and Nebraska/14, using spectra collected at different concentration levels. The trained ML model uses the CNN architecture as shown in Fig 1B in the manuscript. The reported accuracies are spectra-based accuracies, i.e., the percentage of all spectra for a virus sample that are correctly classified as the true label for the virus. The case-based prediction for the virus sample is also reported, which is the majority vote of all the spectra predicted labels.

Testing Virus Strain (400 Raman spectra collected for each strain at each level of dilution)	**Discarding blank spectra from the testing set**		**Undiluted**	**10^{-1}**	**10^{-2}**	**10^{-3}**	**10^{-4}**	**10^{-5}**	**10^{-6}**
Indiana/08 (True label: Flu A, H3N2)	No	Spectra-based accuracy	0.898	0.635	0.510	0.608	0.643	0.515	0.093
	Yes	Spectra-based accuracy (percentage of blank spectra)	0.898 (0% blanks)	0.635 (0% blanks)	0.515 (1% blanks)	0.608 (0.5% blanks)	0.643 (0% blanks)	0.515 (0.25% blanks)	0.949 (90.25% blanks)
	Yes	Case-based prediction	H3N2	H3N2	H3N2	H3N2	H3N2	H3N2	H3N2
Nebraska/14 (True label: Flu A, H1N1)	No	Spectra-based accuracy	0.648	0.855	0.875	0.883	0.755	0.953	0.430
	Yes	Spectra-based accuracy (percentage of blank spectra)	0.648 (0% blanks)	0.855 (0% blanks)	0.888 (1.5% blanks)	0.970 (9% blanks)	0.786 (4% blanks)	0.953 (0.25% blanks)	0.440 (2.25% blanks)
	Yes	Case-based prediction	H1N1	H1N1	H1N1	H1N1	H1N1	H1N1	Flu B (56% spectra predicted as Flu B, 44% predicted as H1N1)