Supplementary Text:
Emergenet: Fast Scalable Pandemic Risk Estimation of Influenza A Strains Collected In Non-human Hosts

Kevin Wu[1], Jin Li[1], Timmy Li[1], Aaron Esser-Kahn[2,3], and Ishanu Chattopadhyay[1,4,5★]

[1]Department of Medicine, University of Chicago, IL, USA
[2]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA
[3]Committee on Immunology, University of Chicago, Chicago, IL, USA
[4]Committee on Genetics, Genomics & Systems Biology, University of Chicago, IL, USA
[5]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, IL, USA

★To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

Supplementary Methods: Notes on q-Distance & Supporting Results

The *q-distance* is a pseudo-metric since distinct sequences can induce the same distributions over each index, and thus evaluate to have a zero distance. This is actually desirable; we do not want our distance to be sensitive to changes that are not biologically relevant. The intuition is that not all sequence variations brought about by substitutions are equally important or likely. Even with no selection pressure, we might still see random variations at an index if such variations do not affect the replicative fitness. Under that scenario, the corresponding Φ_i will predict a flat distribution no matter what the input sequence is, thus contributing nothing to the overall distance. And even if two strains x, y have the same entry at some index i, the remaining residues might induce different distributions Φ_i based on the remote dependencies, $i.e.$, the entries in x_{-i}, y_{-i}. Also, it matters if the sequences come from two different background populations P, Q, $i.e.$, if the induced Qnets Φ^P, Φ^Q are different. Thus, if we construct Qnets for H1N1 Influenza A separately for the collection years 2008 and 2009, then the same exact sequence collected in the respective years might have a non-zero distance between them, reflecting the fact that the background population the sequences arose from are different, inducing possibly different expected mutational tendencies (See SI-Table **??**).

Next, we induce q-distance between a sequence and a population and between two populations.

Definition 1 (Pseudo-metric between populations)**.** *Using the notion of Hausdorff metric between sets:*

$$\forall x \in P, y \in Q,$$
$$\theta(x, Q) = \min_{y \in Q} \theta(x, y) \tag{1}$$
$$\theta(P, Q) = \max\left\{\max_{x \in P} \theta(x, Q), \max_{y \in Q} \theta(y, P)\right\} \tag{2}$$

In-silico Corroboration of Qnet Constraints

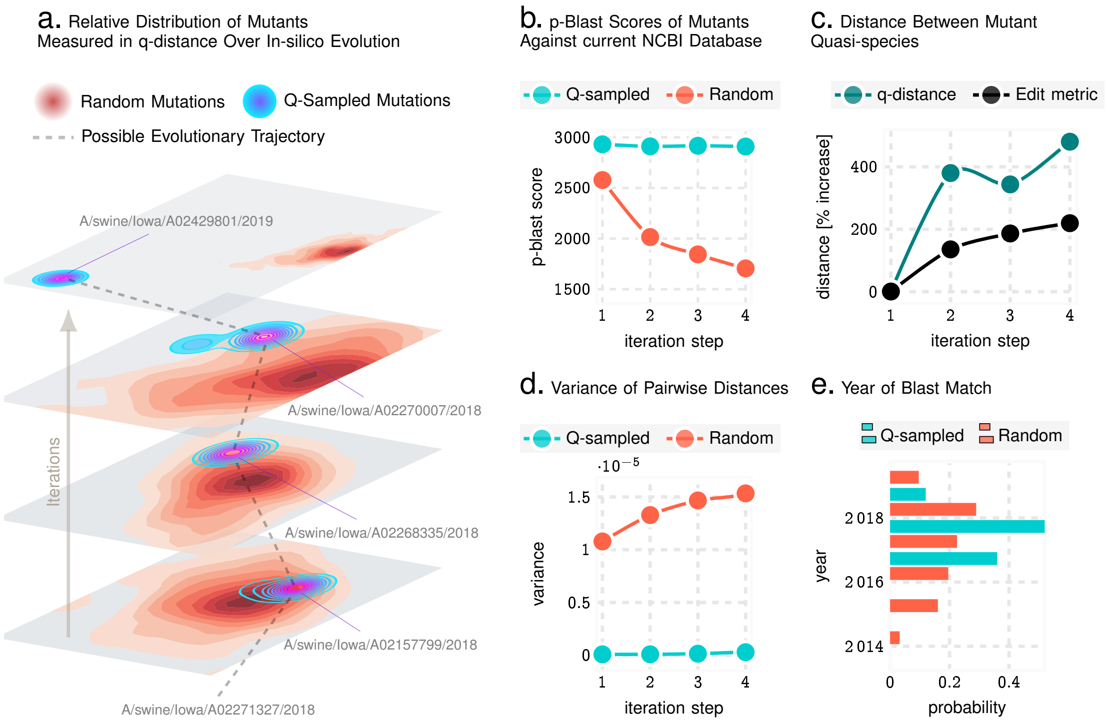
We carry out in-silico experiments to corroborate that the constraints represented within an inferred Qnet are indeed reflective of the biology in play. We compare the results of simulated mutational perturbations to sequences from our databases (for which we have already constructed Qnets), and then use NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify if our perturbed sequences match with existing sequences in the databases (See SI-Fig. 1). We find that in contrast to random variations, which rapidly diverge the trajectories, the Qnet constraints tend to produce smaller variance in the trajectories, maintain a high degree of match as we extend our trajectories, and produces matches closer in time to the collection time of the initial sequence — suggesting that the Qnet does indeed capture realistic constraints.

Multivariate Regression to Identify Factors in Strain Prediction

We investigate the key factors that contribute to our successful prediction of the dominant strain in the next season. We carry out a multivariate regression with data diversity, the complexity of inferred Qnet and the edit distance of the WHO recommendation from the dominant strain as independent variables. Here we define data diversity as the number of clusters we have in the input set of sequences, such that any two sequences five or less mutations apart are in the same cluster. Qnet complexity is measured by the number of decision nodes in the component decision trees of the recursive forest.

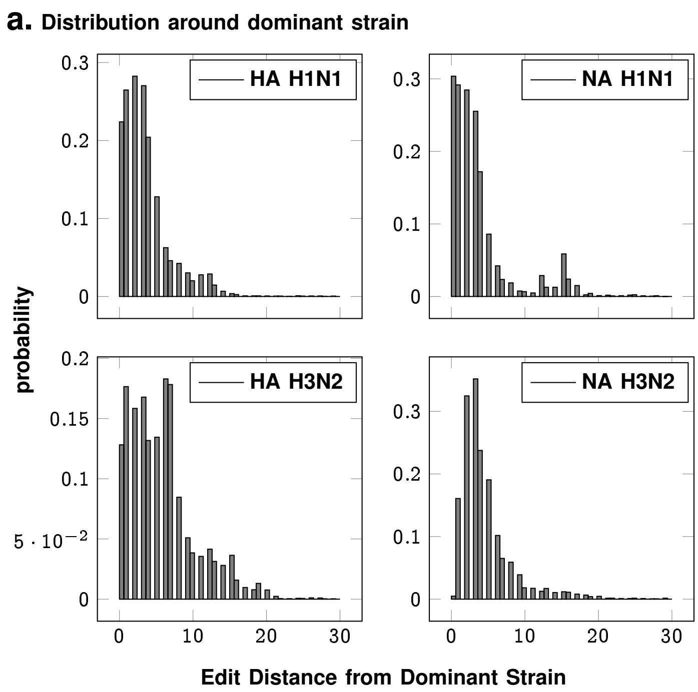
We select several plausible structures of the regression equation, and in each case conclude that data diversity has the most important and statistically significant contribution (See SI-Tab. 12).

REFERENCES



SI Fig. 1. **Q-distance validation in-silico using Influenza A sequences from NCBI database. Panel a** illustrates that the Qnet induced modeling of evolutionary trajectories initiated from known haemagglutinnin (HA) sequences are distinct from random paths in the strain space. In particular, random trajectories have more variance, and more importantly, diverge to different regions of the landscape compared to Qnet predictions. **Panels b-e** show that unconstrained Q-sampling produces sequences maintain a higher degree of similarity to known sequences, as verified by blasting against known HA sequences, have a smaller rate of growth of variance, and produce matches in closer time frames to the initial sequence. **Panel c** shows that this is not due to simply restricting the mutational variations, which increases rapidly in both the Qnet and the classical metric.

SUPPLEMENTARY FIGURES & TABLES



SI Fig. 2. **No. of mutations from the seasonal dominant strain over the years** The quasispecies that circulates each season for each sub-type is tightly distributed around the dominant strain on average.

SI Tab. 1
H1N1 NA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2001-02	A/New Caledonia/20/99	A/New York/447/2001	A/Memphis/15/2000	4	4
2002-03	A/New Caledonia/20/99	A/Paris/0833/2002	A/New York/341/2001	1	5
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-05	A/New Caledonia/20/99	A/Singapore/14/2004	A/New York/223/2003	2	3
2005-06	A/New Caledonia/20/99	A/Taiwan/5524/2005	A/Florida/3e/2004	3	0
2006-07	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-08	A/Solomon Islands/3/2006	A/Tennessee/UR06-0106/2007	A/Sofia/490/2006	9	2
2008-09	A/Brisbane/59/2007	A/Sendai/TU66/2008	A/Maryland/04/2007	0	3
2009-10	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Paris/910/2008	87	87
2010-11	A/California/7/2009	A/Finland/2460N/2010	A/Rome/709/2009	2	9
2011-12	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Oman/SQUH-40/2010	4	2
2012-13	A/California/7/2009	A/Bangalore/697-32/2012	A/Nizhnii Novgorod/CRIE-ZCA/2011	4	0
2013-14	A/California/7/2009	A/Jiangsugusu/SWL1824/2013	A/LongYan/SWL33/2013	5	3
2014-15	A/California/7/2009	A/LongYan/SWL2457/2014	A/Utah/06/2013	9	3
2015-16	A/California/7/2009	A/Michigan/45/2015	A/Maryland/02/2014	14	4
2016-17	A/California/7/2009	A/Mexico/4436/2016	A/India/Pun151245/2015	14	0
2017-18	A/Michigan/45/2015	A/Illinois/37/2017	A/Utah/02/2016	3	3
2018-19	A/Michigan/45/2015	A/Kenya/47/2018	A/Maine/24/2017	4	0
2019-20	A/Brisbane/02/2018	A/Texas/7939/2019	A/Missouri/03/2018	1	0
2020-21	A/Hawaii/70/2019	A/Togo/897/2020	A/Texas/112/2019	0	5
2021-22	A/Victoria/2570/2019	A/Cote_d'Ivoire/3729/2021	A/Togo/0071/2021	1	5
2022-23	-1	-1	A/Lyon/820/2021	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 2
H1N1 NA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2001-02	A/New Caledonia/20/99	A/New York/447/2001	A/Canterbury/37/2000	4	6
2002-03	A/New Caledonia/20/99	A/Paris/0833/2002	A/New York/447/2001	1	5
2003-04	A/New Caledonia/20/99	A/Memphis/5/2003	A/New York/291/2002	3	5
2004-05	A/New Caledonia/20/99	A/Singapore/14/2004	A/Memphis/5/2003	2	3
2005-06	A/New Caledonia/20/99	A/Taiwan/5524/2005	A/Canterbury/106/2004	3	6
2006-07	A/New Caledonia/20/99	A/Massachusetts/08/2006	A/Sofia/361/2005	4	2
2007-08	A/New Caledonia/20/99	A/Tennessee/UR06-0106/2007	A/Thailand/RMSC-UDN-20/2006	4	8
2008-09	A/Solomon Islands/3/2006	A/Sendai/TU66/2008	A/Tennessee/UR06-0151/2007	15	13
2009-10	A/Brisbane/59/2007	A/Thailand/SR08021/2009	A/Nebraska/07/2008	87	87
2010-11	A/California/7/2009	A/Finland/2460N/2010	A/Rome/709/2009	2	9
2011-12	A/California/7/2009	A/Tula/CRIE-GSYu/2011	A/Finland/2460N/2010	4	2
2012-13	A/California/7/2009	A/Bangalore/697-32/2012	A/Tula/CRIE-GSYu/2011	4	0
2013-14	A/California/7/2009	A/Jiangsugusu/SWL1824/2013	A/Oman/SQUH-63/2012	5	4
2014-15	A/California/7/2009	A/LongYan/SWL2457/2014	A/NanPing/SWL1640/2013	9	6
2015-16	A/California/7/2009	A/Michigan/45/2015	A/LongYan/SWL2457/2014	14	5
2016-17	A/California/7/2009	A/Mexico/4436/2016	A/Michigan/45/2015	14	0
2017-18	A/Michigan/45/2015	A/Illinois/37/2017	A/Mexico/4436/2016	3	3
2018-19	A/Michigan/45/2015	A/Kenya/47/2018	A/Kentucky/26/2017	4	2
2019-20	A/Michigan/45/2015	A/Texas/7939/2019	A/Kenya/47/2018	4	0
2020-21	A/Brisbane/02/2018	A/Togo/897/2020	A/Texas/7939/2019	6	5
2021-22	A/Victoria/2570/2019	A/Cote_D'Ivoire/1496/2021	A/NAGASAKI/8/2020	1	6
2022-23	-1	-1	A/Dakar/35/2021	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 3
H3N2 NA Northern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2003-04	A/Moscow/10/99	A/Denmark/107/2003	A/New York/100/2002	13	3
2004-05	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-06	A/California/7/2004	A/Denmark/203/2005	A/Hong Kong/HKU20/2004	4	0
2006-07	A/Wisconsin/67/2005	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	1	1
2007-08	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Baden-Wuerttemberg/17/2006	8	7
2008-09	A/Brisbane/10/2007	A/Missouri/05/2008	A/Washington/01/2007	3	2
2009-10	A/Brisbane/10/2007	A/Oklahoma/09/2009	A/Wisconsin/24/2008	3	1
2010-11	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-12	A/Perth/16/2009	A/Texas/14/2011	A/California/14/2010	3	2
2012-13	A/Victoria/361/2011	A/New York/02/2012	A/Singapore/C2011.493/2011	4	1
2013-14	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/01/2012	3	1
2014-15	A/Texas/50/2012	A/Tehran/69634/2014	A/Boston/DOA2-176/2013	3	1
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Thailand/CU-B10520/2014	3	0
2016-17	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Delaware/02/2015	7	2
2017-18	A/Hong Kong/4801/2014	A/Texas/277/2017	A/New York/03/2016	8	0
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Colorado/11/2017	4	3
2019-20	A/Kansas/14/2017	A/Washington/9757/2019	A/Guangxi-Fangcheng/54/2019	3	11
2020-21	A/Hong Kong/2671/2019	A/Bangladesh/1004005/2020	A/Maryland/02/2019	3	13
2021-22	A/Cambodia/e0826360/2020	A/Stockholm/10/2022	A/Bangladesh/1916/2020	2	2
2022-23	-1	-1	A/Iowa/20/2022	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 4
H3N2 NA Southern Hemisphere

Year	WHO Recommendation	Dominant Strain	Qnet Recommendation	WHO Error	Qnet Error
2003-04	A/Moscow/10/99	A/Denmark/107/2003	A/New York/101/2002	13	3
2004-05	A/Fujian/411/2002	A/Hyogo/36/2004	A/New York/20/2003	3	16
2005-06	A/Wellington/1/2004	A/Denmark/203/2005	A/Wellington/1/2004	2	2
2006-07	A/California/7/2004	A/Berlin/32/2006	A/Mexico/InDRE2227/2005	3	1
2007-08	A/Wisconsin/67/2005	A/Brazil/80/2007	A/Ohio/06/2006	8	10
2008-09	A/Brisbane/10/2007	A/Missouri/05/2008	A/Brazil/80/2007	3	2
2009-10	A/Brisbane/10/2007	A/Oklahoma/09/2009	A/Wisconsin/24/2008	3	1
2010-11	A/Perth/16/2009	A/California/17/2010	A/New York/70/2009	2	3
2011-12	A/Perth/16/2009	A/Texas/14/2011	A/Virginia/05/2010	3	2
2012-13	A/Perth/16/2009	A/New York/02/2012	A/Texas/14/2011	4	1
2013-14	A/Victoria/361/2011	A/Michigan/02/2013	A/New York/02/2012	3	3
2014-15	A/Texas/50/2012	A/Tehran/69634/2014	A/Michigan/02/2013	3	1
2015-16	A/Switzerland/9715293/2013	A/Parma/471/2015	A/Tehran/69634/2014	3	2
2016-17	A/Hong Kong/4801/2014	A/North Carolina/62/2016	A/Parma/471/2015	7	2
2017-18	A/Hong Kong/4801/2014	A/Texas/277/2017	A/Guangdong/264/2016	8	0
2018-19	A/Singapore/INFIMH-16-0019/2016	A/Japan/NHRC_FDX70352/2018	A/Texas/277/2017	4	3
2019-20	A/Switzerland/8060/2017	A/Washington/9757/2019	A/Pennsylvania/317/2018	10	10
2020-21	A/South Australia/34/2019	A/Bangladesh/1004005/2020	A/Washington/9757/2019	1	13
2021-22	A/Hong Kong/2671/2019	A/India/PUN-NIV301718/2021	A/India/PUN-NIV301132/2021	6	4
2022-23	-1	-1	A/Michigan/UOM10045036720/2022	-1	-1

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 5
H1N1 NA NORTHERN HEMISPHERE (MULTI-CLUSTER)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2001-02	A/New Caledonia/20/99	4	1	6	A/New South Wales/26/2000	A/Canterbury/37/2000
2002-03	A/New Caledonia/20/99	1	0	5	A/Wellington/1/2001	A/New York/447/2001
2003-04	A/New Caledonia/20/99	3	2	8	A/Paris/0833/2002	A/Taiwan/141/2002
2004-05	A/New Caledonia/20/99	2	3	4	A/Memphis/5/2003	A/Hanoi/1004/2003
2005-06	A/New Caledonia/20/99	3	0	1	A/Denmark/130/2004	A/Paris/650/2004
2006-07	A/New Caledonia/20/99	4	2	8	A/Sofia/361/2005	A/Wellington/11/2005
2007-08	A/Solomon Islands/3/2006	9	4	8	A/Sofia/246/2006	A/New York/8/2006
2008-09	A/Brisbane/59/2007	0	13	19	A/Tennessee/UR06-0151/2007	A/Ohio/UR06-0178/2007
2009-10	A/Brisbane/59/2007	87	88	90	A/Sendai/TU66/2008	A/Japan/618/2008
2010-11	A/California/7/2009	2	1	6	A/South Carolina/WRAIR1645P/2009	A/Wisconsin/629-D00809/2009
2011-12	A/California/7/2009	4	1	3	A/England/21680633/2010	A/Hangzhou/178/2010
2012-13	A/California/7/2009	4	1	22	A/Joshkar-Ola/CRIE-BLP/2011	A/Rio Grande do Sul/578/2011
2013-14	A/California/7/2009	5	4	13	A/Thailand/MR10580/2012	A/Mexico/INMEGEN-INER 15/2012
2014-15	A/California/7/2009	9	3	7	A/Minnesota/02/2013	A/Helsinki/430/2013
2015-16	A/California/7/2009	14	4	7	A/Helsinki/808M/2014	A/Virginia/NHRC430739/2014
2016-17	A/California/7/2009	14	0	3	A/Michigan/45/2015	A/Colorado/30/2015
2017-18	A/Michigan/45/2015	3	3	8	A/Mexico/4436/2016	A/Arizona/03/2016
2018-19	A/Michigan/45/2015	4	0	4	A/California/NHRC_QV11073/2017	A/Minnesota/35/2017
2019-20	A/Brisbane/02/2018	1	0	2	A/Kenya/47/2018	A/Colorado/7682/2018
2020-21	A/Hawaii/70/2019	0	3	8	A/California/NHRC-OID_BOX-ILI-0012/2019	A/Indiana/30/2019
2021-22	A/Victoria/2570/2019	1	5	51	A/Togo/0071/2021	A/Yunnan-Mengzi/1462/2020
2022-23	-1	-1	-1	-1	A/Netherlands/10646/2022	A/Sydney/234/2022

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 6
H1N1 NA SOUTHERN HEMISPHERE (MULTI-CLUSTER)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2001-02	A/New Caledonia/20/99	4	1	6	A/New South Wales/26/2000	A/Canterbury/37/2000
2002-03	A/New Caledonia/20/99	1	0	5	A/Wellington/1/2001	A/New York/447/2001
2003-04	A/New Caledonia/20/99	3	2	8	A/Paris/0833/2002	A/Taiwan/141/2002
2004-05	A/New Caledonia/20/99	2	3	4	A/Memphis/5/2003	A/Hanoi/1004/2003
2005-06	A/New Caledonia/20/99	3	0	1	A/Denmark/130/2004	A/Paris/650/2004
2006-07	A/New Caledonia/20/99	4	2	8	A/Sofia/361/2005	A/Wellington/11/2005
2007-08	A/New Caledonia/20/99	4	4	8	A/Sofia/246/2006	A/New York/8/2006
2008-09	A/Solomon Islands/3/2006	15	13	19	A/Tennessee/UR06-0151/2007	A/Ohio/UR06-0178/2007
2009-10	A/Brisbane/59/2007	87	88	90	A/Sendai/TU66/2008	A/Japan/618/2008
2010-11	A/California/7/2009	2	1	6	A/South Carolina/WRAIR1645P/2009	A/Wisconsin/629-D00809/2009
2011-12	A/California/7/2009	4	1	3	A/England/21680633/2010	A/Hangzhou/178/2010
2012-13	A/California/7/2009	4	1	22	A/Joshkar-Ola/CRIE-BLP/2011	A/Rio Grande do Sul/578/2011
2013-14	A/California/7/2009	5	4	13	A/Thailand/MR10580/2012	A/Mexico/INMEGEN-INER 15/2012
2014-15	A/California/7/2009	9	3	7	A/Minnesota/02/2013	A/Helsinki/430/2013
2015-16	A/California/7/2009	14	4	7	A/Helsinki/808M/2014	A/Virginia/NHRC430739/2014
2016-17	A/California/7/2009	14	0	3	A/Michigan/45/2015	A/Colorado/30/2015
2017-18	A/Michigan/45/2015	3	3	8	A/Mexico/4436/2016	A/Arizona/03/2016
2018-19	A/Michigan/45/2015	4	0	4	A/California/NHRC_QV11073/2017	A/Minnesota/35/2017
2019-20	A/Michigan/45/2015	4	0	2	A/Kenya/47/2018	A/Colorado/7682/2018
2020-21	A/Brisbane/02/2018	5	2	7	A/California/NHRC-OID_BOX-ILI-0012/2019	A/Indiana/30/2019
2021-22	A/Victoria/2570/2019	1	7	58	A/Togo/0155/2021	A/Shandong/00204/2021
2022-23	-1	-1	-1	-1	A/Switzerland/86136/2022	A/Wisconsin/04/2021

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 7
H3N2 NA NORTHERN HEMISPHERE (MULTI-CLUSTER)

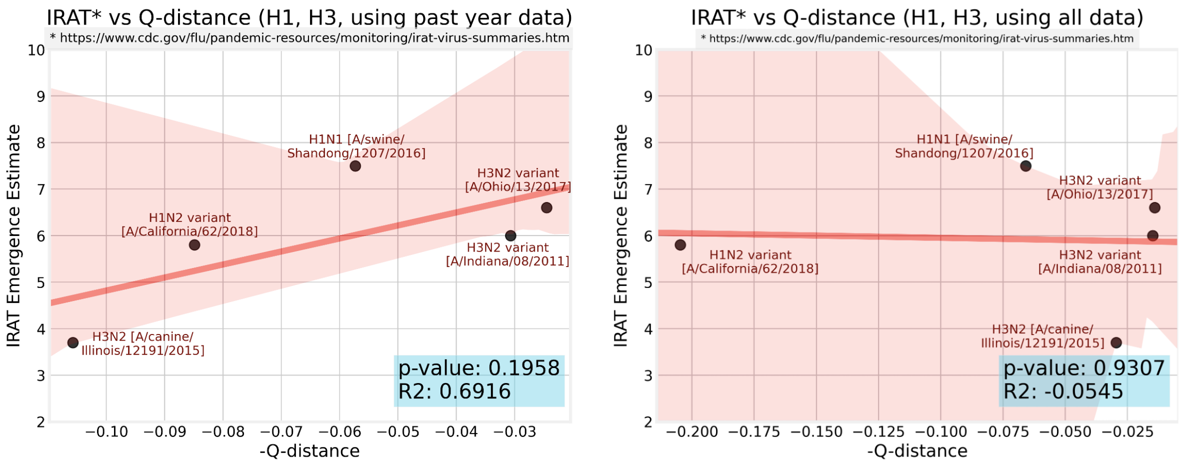
Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2003-04	A/Moscow/10/99	13	4	5	A/Auckland/612/2002	A/New York/87/2002
2004-05	A/Fujian/411/2002	3	16	18	A/New York/20/2003	A/New York/12/2003
2005-06	A/California/7/2004	4	1	7	A/New York/358/2004	A/Singapore/36/2004
2006-07	A/Wisconsin/67/2005	1	3	8	A/Macau/557/2005	A/Hong Kong/HKU53/2005
2007-08	A/Wisconsin/67/2005	8	0	10	A/Wisconsin/42/2006	A/Wisconsin/44/2006
2008-09	A/Brisbane/10/2007	3	4	10	A/Missouri/06/2007	A/Japan/72/2007
2009-10	A/Brisbane/10/2007	3	1	7	A/Wisconsin/24/2008	A/Mississippi/UR07-0042/2008
2010-11	A/Perth/16/2009	2	3	8	A/New York/70/2009	A/Japan/883/2009
2011-12	A/Perth/16/2009	3	2	2	A/California/19/2010	A/Virginia/05/2010
2012-13	A/Victoria/361/2011	4	1	12	A/Texas/14/2011	A/Singapore/GP1684/2011
2013-14	A/Victoria/361/2011	3	1	5	A/Idaho/38/2012	A/Pavia/135/2012
2014-15	A/Texas/50/2012	3	1	1	A/Nevada/05/2013	A/Michigan/02/2013
2015-16	A/Switzerland/9715293/2013	3	0	4	A/Nicaragua/6866_14/2014	A/Iran/91244/2014
2016-17	A/Hong Kong/4801/2014	7	1	25	A/New Jersey/13/2015	A/California/NHRC_BRD41056N/2015
2017-18	A/Hong Kong/4801/2014	9	1	4	A/Guangdong/264/2016	A/Victoria/668/2016
2018-19	A/Singapore/INFIMH-16-0019/2016	3	2	4	A/Netherlands/3530/2017	A/Washington/17/2017
2019-20	A/Kansas/14/2017	3	4	10	A/England/538/2018	A/California/BRD12490N/2018
2020-21	A/Hong Kong/2671/2019	3	1	13	A/England/9738/2019	A/Washington/9757/2019
2021-22	A/Cambodia/e0826360/2020	2	3	7	A/Laos/527/2021	A/Michigan/UOM10045655748/2020
2022-23	-1	-1	-1	-1	A/Maine/02/2022	A/Michigan/UOM10042819294/2021

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric

SI Tab. 8
H3N2 NA SOUTHERN HEMISPHERE (MULTI-CLUSTER)

Year	WHO Recommendation	WHO Error	Qnet Error 1	Qnet Error 2	Qnet Recommendation 1	Qnet Recommendation 2
2003-04	A/Moscow/10/99	13	4	5	A/Auckland/612/2002	A/New York/87/2002
2004-05	A/Fujian/411/2002	3	16	18	A/New York/20/2003	A/New York/12/2003
2005-06	A/Wellington/1/2004	2	1	7	A/New York/358/2004	A/Singapore/36/2004
2006-07	A/California/7/2004	3	3	8	A/Macau/557/2005	A/Hong Kong/HKU53/2005
2007-08	A/Wisconsin/67/2005	8	0	10	A/Wisconsin/42/2006	A/Wisconsin/44/2006
2008-09	A/Brisbane/10/2007	3	4	10	A/Missouri/06/2007	A/Japan/72/2007
2009-10	A/Brisbane/10/2007	3	1	7	A/Wisconsin/24/2008	A/Mississippi/UR07-0042/2008
2010-11	A/Perth/16/2009	2	3	8	A/New York/70/2009	A/Japan/883/2009
2011-12	A/Perth/16/2009	3	2	2	A/California/19/2010	A/Virginia/05/2010
2012-13	A/Perth/16/2009	4	1	12	A/Texas/14/2011	A/Singapore/GP1684/2011
2013-14	A/Victoria/361/2011	3	1	5	A/Idaho/38/2012	A/Pavia/135/2012
2014-15	A/Texas/50/2012	3	1	1	A/Nevada/05/2013	A/Michigan/02/2013
2015-16	A/Switzerland/9715293/2013	3	0	4	A/Nicaragua/6866_14/2014	A/Iran/91244/2014
2016-17	A/Hong Kong/4801/2014	7	1	25	A/New Jersey/13/2015	A/California/NHRC_BRD41056N/2015
2017-18	A/Hong Kong/4801/2014	9	1	4	A/Guangdong/264/2016	A/Victoria/668/2016
2018-19	A/Singapore/INFIMH-16-0019/2016	3	2	4	A/Netherlands/3530/2017	A/Washington/17/2017
2019-20	A/Switzerland/8060/2017	10	4	10	A/England/538/2018	A/California/BRD12490N/2018
2020-21	A/South Australia/34/2019	1	1	13	A/England/9738/2019	A/Washington/9757/2019
2021-22	A/Hong Kong/2671/2019	6	1	49	A/Darwin/11/2021	A/Hawaii/28/2020
2022-23	-1	-1	-1	-1	A/Congo/313/2021	A/Texas/12723/2022

* Dominant strain is calculated as the one closest to the centroid in the strain space that year in the edit distance metric



SI Fig. 3. **IRAT vs. Q-distance relationship for H1- and H3- sub-types, using past year data vs. using all data.** On the result when computing average q-distance between the target strain and the circulating human strains from the past year, and on the right is the result when using all available human strains of that sub-type. Evidently, the former has a much higher correlation, since a strain being “close” to humans at some point does not necessarily mean being close now.

SI Tab. 9
INFLUENZA A STRAINS EVALUATED BY IRAT AND CORRESPONDING QNET COMPUTED RISK SCORES

Influenza Virus	Subype	IRAT Date	IRAT Emergence Score	IRAT Impact Score	HA Qnet Sample	NA Qnet Sample	HA Avg. Q-dist.	NA Avg. Q-dist.	Geom. Mean	Qnet Emergence Score	Qnet Impact Score
A/swine/Shandong/1207/2016	H1N1	Jul 2020	7.5	6.9	1000	1000	0.0941	0.0205	0.0440	6.0	6.2
A/Ohio/13/2017	H3N2	Jul 2019	6.6	5.8	1000	1000	0.0184	0.0306	0.0238	6.3	6.2
A/Hong Kong/125/2017	H7N9	May 2017	6.5	7.5	437	437	0.0296	0.0058	0.0131	6.6	6.5
A/Shanghai/02/2013	H7N9	Apr 2016	6.4	7.2	178	178	0.0055	0.0036	0.0044	6.7	6.6
A/Anhui-Lujiang/39/2018	H9N2	Jul 2019	6.2	5.9	31	30	0.0290	0.1681	0.0698	5.2	5.0
A/Indiana/08/2011	H3N2	Dec 2012	6.0	4.5	1000	1000	0.0523	0.0091	0.0218	6.4	6.5
A/California/62/2018	H1N2	Jul 2019	5.8	5.7	55	55	0.1089	0.0610	0.0815	5.4	5.5
A/Bangladesh/0994/2011***	H9N2	Feb 2014	5.6	5.4	-1	-1	0.2078	0.1823	0.1947	4.3	4.9
A/Sichuan/06681/2021	H5N6	Oct 2021	5.3	6.3	45	45	0.3616	0.0518	0.1369	5.2	6.4
A/Vietnam/1203/2004	H5N1	Nov 2011	5.2	6.6	258	246	0.1673	0.0111	0.0430	6.2	6.7
A/Yunnan/14564/2015**	H5N6	Apr 2016	5.0	6.6	344	331	0.3482	0.2987	0.3225	4.9	6.5
A/Astrakhan/3212/2020**	H5N8	Mar 2021	4.6	5.2	381	365	0.1603	0.3472	0.2359	3.9	4.4
A/Netherlands/219/2003	H7N7	Jun 2012	4.6	5.8	46	46	0.2757	0.3521	0.3115	4.6	5.8
A/American wigeon/South Carolina/AH0195145/2021	H5N1	Mar 2022	4.4	5.1	335	323	0.1722	0.5114	0.2967	4.0	4.7
A/Jiangxi-Donghu/346/2013***	H10N8	Feb 2014	4.3	6.0	-1	-1	0.2088	0.2101	0.2094	4.3	4.8
A/gyrfalcon/Washington/ 41088/2014**	H5N8	Mar 2015	4.2	4.6	341	328	0.1532	0.3424	0.2290	3.9	4.3
A/Northern pintail/ Washington/40964/2014**	H5N2	Mar 2015	3.8	4.1	341	328	0.1529	0.3799	0.2410	3.9	4.3
A/canine/Illinois/12191/2015	H3N2	Jun 2016	3.7	3.7	1000	1000	0.0607	0.1509	0.0957	4.9	4.8
A/American green-winged teal /Washington/1957050/2014	H5N1	Mar 2015	3.6	4.1	326	314	0.1911	0.4482	0.2927	4.1	4.9
A/turkey/Indiana/1573-2/2016**	H7N8	Jul 2017	3.4	3.9	495	494	0.1130	0.7738	0.2957	3.4	3.9
A/chicken/Tennessee/17-007431-3/2017	H7N9	Oct 2017	3.1	3.5	496	495	0.1027	0.2569	0.1624	4.1	4.2
A/chicken/Tennessee/17-007147-2/2017	H7N9	Oct 2017	2.8	3.5	496	495	0.2095	0.2541	0.2307	4.2	4.8
A/duck/New York/1996 *	H1N1	Nov 2011	2.3	2.4	1000	1000	-1	-1	-1	-1	-1

* HA strain is not available for A/duck/New York/1996, so this strain is omitted.
** Could not construct a Qnet of human sequence data available for that virus sub-type (less than 30 strains), so we constructed a Qnet using all human strains that match the HA sub-type, i.e. H5NX for H5N6.
*** These strains did not have enough human sequence data to generate a Qnet, even when only considering the HA sub-type. Thus, we estimated the risk score using every Qnet from the other IRAT strains, and took the average among NA and HA. Finally, we took the geometric mean of the resulting NA and HA averages.

SI Tab. 10
INFLUENZA A STRAINS EVALUATED BY IRAT AND CORRESPONDING QNET COMPUTED CURRENT RISK SCORES

Influenza Virus	Subype	IRAT Date	IRAT Emer-gence Score	IRAT Impact Score	HA Qnet Sample	NA Qnet Sample	HA Avg. Q-dist.	NA Avg. Q-dist.	Geom. Mean	Qnet Emer-gence Score	Qnet Impact Score
A/swine/Shandong/1207/2016	H1N1	Jul 2020	7.5	6.9	1000	1000	0.0599	0.0417	0.0500	5.8	5.8
A/Ohio/13/2017	H3N2	Jul 2019	6.6	5.8	1000	1000	0.0091	0.0692	0.0251	6.2	6.0
A/Hong Kong/125/2017	H7N9	May 2017	6.5	7.5	1000	1000	0.0092	0.0046	0.0065	6.7	6.6
A/Shanghai/02/2013	H7N9	Apr 2016	6.4	7.2	1000	1000	0.0031	0.0044	0.0037	6.8	6.6
A/Anhui-Lujiang/39/2018	H9N2	Jul 2019	6.2	5.9	58	58	0.0157	0.0467	0.0271	6.2	6.0
A/Indiana/08/2011	H3N2	Dec 2012	6.0	4.5	1000	1000	0.0176	0.0184	0.0180	6.4	6.3
A/California/62/2018	H1N2	Jul 2019	5.8	5.7	37	37	0.2038	0.0477	0.0986	5.3	5.9
A/Bangladesh/0994/2011	H9N2	Feb 2014	5.6	5.4	58	58	0.0473	0.4654	0.1484	3.8	3.6
A/Sichuan/06681/2021	H5N6	Oct 2021	5.3	6.3	46	46	0.3443	0.0600	0.1437	5.1	6.2
A/Vietnam/1203/2004	H5N1	Nov 2011	5.2	6.6	48	45	0.1323	0.0411	0.0738	5.6	5.8
A/Yunnan/14564/2015	H5N6	Apr 2016	5.0	6.6	46	46	0.2187	0.0415	0.0953	5.4	6.0
A/Astrakhan/3212/2020	H5N8	Mar 2021	4.6	5.2	95	92	0.2366	0.5451	0.3591	4.8	6.1
A/Netherlands/219/2003	H7N7	Jun 2012	4.6	5.8	1000	1000	0.1658	0.4596	0.2760	3.9	4.5
A/American wigeon/South Carolina/AH0195145/2021	H5N1	Mar 2022	4.4	5.1	48	45	0.2355	0.3135	0.2717	4.3	5.2
A/Jiangxi-Donghu/346/2013**	H10N8	Feb 2014	4.3	6.0	-1	-1	0.2097	0.2299	0.2196	4.2	4.8
A/gyrfalcon/Washington/ 41088/2014	H5N8	Mar 2015	4.2	4.6	95	92	0.2387	0.5438	0.3603	4.8	6.1
A/Northern pintail /Washington/40964/2014	H5N2	Mar 2015	3.8	4.1	95	92	0.2327	0.5099	0.3445	4.6	5.8
A/canine/Illinois/12191/2015	H3N2	Jun 2016	3.7	3.7	1000	1000	0.0179	0.0374	0.0259	6.2	6.1
A/American green-winged teal /Washington/1957050/2014	H5N1	Mar 2015	3.6	4.1	48	45	0.2352	0.3067	0.2686	4.3	5.1
A/turkey/Indiana/1573-2/2016	H7N8	Jul 2017	3.4	3.9	1000	1000	0.0438	0.4165	0.1351	4.0	3.8
A/chicken/Tennessee/17-007431-3/2017	H7N9	Oct 2017	3.1	3.5	1000	1000	0.0335	0.5127	0.1310	3.8	3.6
A/chicken/Tennessee/17-007147-2/2017	H7N9	Oct 2017	2.8	3.5	1000	1000	0.0839	0.5127	0.2075	3.5	3.6
A/duck/New York/1996*	H1N1	Nov 2011	2.3	2.4	1000	1000	-1	-1	-1	-1	-1

*This table contains Qnet scores for IRAT computed using current sequence data, thereby computing the current risk of these strains. -1 indicates missing data, either from lack of human sequence data available for that virus sub-type (less than 30 strains) or missing IRAT sequence data (in the case of A/duck/New York/1996)

SI Tab. 11
GENERAL LINEAR MODEL FOR EVALUATING EFFECT OF DATA DIVERSITY ON QNET PERFORMANCE

Variable Name	Description
qnet_complexity	Cumulative number of nodes in all predictors in the corresponding Qnet
data_diversity	Number of clusters in set of input sequence where each sequence in a specific cluster is separated by at least 5 mutations from sequences not in the cluster
ldistance_WHO	Deviation of WHO predicted strain from the dominant strain

```
model:dev ~ qnet_complexity + data_diversity + qnet_complexity * data_diversity + ldistance_WHO
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                    dev   No. Observations:                  235
Model:                            GLM   Df Residuals:                      230
Model Family:                Gaussian   Df Model:                            4
Link Function:               identity   Scale:                          23.214
Method:                          IRLS   Log-Likelihood:                -700.43
Date:                Thu, 11 Jun 2020   Deviance:                       5339.2
Time:                        16:45:46   Pearson chi2:                 5.34e+03
No. Iterations:                     3   Covariance Type:             nonrobust
===================================================================================================
                                      coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                          -0.1116      1.090     -0.102      0.918      -2.248       2.025
qnet_complexity                     0.0005      0.000      1.075      0.282      -0.000       0.001
data_diversity                      0.3197      0.126      2.531      0.011       0.072       0.567
qnet_complexity:data_diversity -6.932e-05   5.01e-05     -1.383      0.167      -0.000    2.89e-05
ldistance_WHO                      -0.0348      0.035     -1.007      0.314      -0.102       0.033
===================================================================================================
```

```
model:dev ~ qnet_complexity + data_diversity + ldistance_WHO
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                    dev   No. Observations:                  235
Model:                            GLM   Df Residuals:                      231
Model Family:                Gaussian   Df Model:                            3
Link Function:               identity   Scale:                          23.306
Method:                          IRLS   Log-Likelihood:                -701.41
Date:                Thu, 11 Jun 2020   Deviance:                       5383.6
Time:                        16:45:47   Pearson chi2:                 5.38e+03
No. Iterations:                     3   Covariance Type:             nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           1.0841      0.665      1.630      0.103      -0.219       2.387
qnet_complexity  -4.12e-05      0.000     -0.156      0.876      -0.001       0.000
data_diversity      0.1788      0.075      2.392      0.017       0.032       0.325
ldistance_WHO      -0.0695      0.024     -2.930      0.003      -0.116      -0.023
===================================================================================
```

SI Tab. 12

GENERAL LINEAR MODEL EVALUATING QNET EMERGENCE RISK PREDICTIONS AGAINST IRAT ESTIMATES

```
Model: IRAT_Emergence_Score ~ Geometric_Mean
================================================================================
Dep. Variable:     IRAT_Emergence_Score   No. Observations:                   22
Model:                              GLM   Df Residuals:                       20
Model Family:                  Gaussian   Df Model:                            1
Link Function:                 identity   Scale:                         0.86853
Method:                            IRLS   Log-Likelihood:                -28.618
Date:                  Tue, 25 Oct 2022   Deviance:                       17.371
Time:                          00:58:27   Pearson chi2:                     17.4
No. Iterations:                       3   Pseudo R-squ. (CS):             0.5919
Covariance Type:              nonrobust
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          6.2467      0.356     17.529      0.000       5.548       6.945
Geometric_Mean    -8.1063      1.830     -4.429      0.000     -11.693      -4.519
==================================================================================
```

```
Model: IRAT_Emergence_Score ~ Geometric_Mean + HA_Avg_Qdist*NA_Avg_Qdist
================================================================================
Dep. Variable:     IRAT_Emergence_Score   No. Observations:                   22
Model:                              GLM   Df Residuals:                       17
Model Family:                  Gaussian   Df Model:                            4
Link Function:                 identity   Scale:                         0.69369
Method:                            IRLS   Log-Likelihood:                -24.357
Date:                  Tue, 25 Oct 2022   Deviance:                       11.793
Time:                          00:58:59   Pearson chi2:                     11.8
No. Iterations:                       3   Pseudo R-squ. (CS):             0.7797
Covariance Type:              nonrobust
=============================================================================================
                                coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------------------
Intercept                     6.8403      0.442     15.459      0.000       5.973       7.708
Geometric_Mean              -23.7466      9.674     -2.455      0.014     -42.707      -4.786
HA_Avg_Qdist                  1.9097      3.979      0.480      0.631      -5.889       9.708
NA_Avg_Qdist                 -1.8133      2.826     -0.642      0.521      -7.353       3.726
HA_Avg_Qdist:NA_Avg_Qdist    54.2280     21.474      2.525      0.012      12.139      96.317
=============================================================================================
```

SI Tab. 13

NUMBERING CONVERSION TO PDM09 AND H3 SCHEMES

Query	H1N1pdm	H3	Query	H1N1pdm	H3	Query	H1N1pdm	H3	Query	H1N1pdm	H3
1	-	-	77	60	69	157	140	143	-	-	-
2	-	-	78	61	70	158	141	144	-	-	-
3	-	-	79	62	71	159	142	145	-	-	-
4	-	-	80	63	72	160	143	146	238	221	224
5	-	-	81	64	73	161	144	147	239	222	225
6	-	-	82	65	74	162	145	148	240	223	226
7	-	-	83	66	75	163	146	149	241	224	227
8	-	-	84	67	76	164	147	150	242	225	228
9	-	-	85	68	77	165	148	151	243	226	229
10	-	-	86	69	78	166	149	152	244	227	230
11	-	-	87	70	79	167	150	153	245	228	231
12	-	-	88	71	80	168	151	154	246	229	232
13	-	-	89	72	81	169	152	155	247	230	233
14	-	-	90	73	82	170	153	156	248	231	234
15	-	-	91	74	-	171	154	157	249	232	235
16	-	-	92	75	83	172	155	158	250	233	236
17	-	-	93	76	84	-	-	-	251	234	237
-	-	1	94	77	85	-	-	-	252	235	238
-	-	2	95	78	86	-	-	-	253	236	239
-	-	3	96	79	87	-	-	-	254	237	240
-	-	4	97	80	88	173	156	159	255	238	241
-	-	5	98	81	89	174	157	160	256	239	242
-	-	6	99	82	90	175	158	161	257	240	243
-	-	7	100	83	91	176	159	162	258	241	244
-	-	8	101	84	92	177	160	163	259	242	245
-	-	9	102	85	-	178	161	164	260	243	246
-	-	10	103	86	93	179	162	165	261	244	247
18	1	11	104	87	94	180	163	166	262	245	248
19	2	12	105	88	95	181	164	167	263	246	249
20	3	13	106	89	96	182	165	168	264	247	250
21	4	14	107	90	97	183	166	169	265	248	251
22	5	15	108	91	98	184	167	170	266	249	252
23	6	16	109	92	99	-	-	-	267	250	253
24	7	17	110	93	100	185	168	171	268	251	254
25	8	18	111	94	101	186	169	172	269	252	255
26	9	19	112	95	102	187	170	173	270	253	256
27	10	20	-	-	-	-	-	-	271	254	257
28	11	21	-	-	-	188	171	174	272	255	258
29	12	22	113	96	103	189	172	175	273	256	259
30	13	23	114	97	104	190	173	176	274	257	260
31	14	24	115	98	105	191	174	177	275	258	261
32	15	25	116	99	106	192	175	178	276	259	262
33	16	26	117	100	107	193	176	179	-	-	-
34	17	27	118	101	108	194	177	180	-	-	-
35	18	28	119	102	109	195	178	181	-	-	-
36	19	29	120	103	110	196	179	182	-	-	-
37	20	30	121	104	111	197	180	183	-	-	-
38	21	31	122	105	112	198	181	184	-	-	-
39	22	32	123	106	113	199	182	185	-	-	-
40	23	33	124	107	114	200	183	186	-	-	-
41	24	34	125	108	115	201	184	187	-	-	-
42	25	35	126	109	116	202	185	188	277	260	-
43	26	36	127	110	117	203	186	189	278	261	263
44	27	37	128	111	118	204	187	190	279	262	264
45	28	38	129	112	119	205	188	191	280	263	265
46	29	39	130	113	120	206	189	192	281	264	266
47	30	40	131	114	121	207	190	193	282	265	267
48	31	41	132	115	122	208	191	194	283	266	268
49	32	42	133	116	123	209	192	195	284	267	269
50	33	43	-	-	-	210	193	196	285	268	270
51	34	44	-	-	-	211	194	197	286	269	271
52	35	45	134	117	124	212	195	198	287	270	272
53	36	46	135	118	125	213	196	199	288	271	273
54	37	47	136	119	-	-	-	-	289	272	274
55	38	48	137	120	-	214	197	200	290	273	275
56	39	49	138	121	-	215	198	201	291	274	276
57	40	50	139	122	126	216	199	202	292	275	277
58	41	51	140	123	127	217	200	203	293	276	278
59	42	52	141	124	128	218	201	204	294	277	279
60	43	53	-	-	-	219	202	205	295	278	280
61	44	54	-	-	-	220	203	206	296	279	281
62	45	-	-	-	-	221	204	207	297	280	282
63	46	55	-	-	-	222	205	208	298	281	283
64	47	56	-	-	-	223	206	209	299	282	284
65	48	57	142	125	129	224	207	210	300	283	285
66	49	58	143	126	130	225	208	211	-	-	-
67	50	59	144	127	131	226	209	212	301	284	286
68	51	60	145	128	132	227	210	213	302	285	287
-	-	-	146	129	133	228	211	214	303	286	288
-	-	-	147	130	-	229	212	215	304	287	289
-	-	-	148	131	134	230	213	216	305	288	290
-	-	-	149	132	135	231	214	217	306	289	291
-	-	-	150	133	136	232	215	218	307	290	292
69	52	61	151	134	137	233	216	219	308	291	293
70	53	62	152	135	138	234	217	220	309	292	294
71	54	63	153	136	139	235	218	221	310	293	295
72	55	64	154	137	140	236	219	222	311	294	296
73	56	65	155	138	141	237	220	223	-	-	-
74	57	66	-	-	-	-	-	-	312	295	297
75	58	67	156	139	142	-	-	-	313	296	298