Supplementary Material

1 EVALUATION METRICS

Equations of five evaluation metrics used in the study are presented as follows, respectively.

$$\text{Accuracy } = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Precision } = \frac{TP}{TP+FP}$$

$$\text{Recall } = \frac{TP}{TP+FN}$$

$$\text{F-Score } = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{MCC (Matthews correlation coefficient) } = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP, FP, TN, FN are calculated from the results of the binary classification task:

- TP: Number of **True Positive** results. In our task, it means the number of samples which will mutate in the future and are correctly predicted by our model.
- FP: Number of **False Positive** results. In our task, it means the number of samples which will not mutate in the future but are wrongly predicted by our model.
- TN: Number of **True Negative** results. In our task, it means the number of samples which will not mutate in the future and are correctly predicted by our model.
- FN: Number of **False Negative** results. In our task, it means the number of samples which will mutate in the future but are wrongly predicted by our model.

2 AN EXAMPLE OF HISTORICAL SEQUENCE DATA GENERATED BY PT-BASED SAMPLING METHOD

An example of the generated historical sequences is shown in Table S1. The PT-based sampling method will generate historical sequences with different lengths. For the purpose of example, we present the produced historical sequences with a length of 5 with the consideration of illustration convenience. Each row in the table refers to a historical sequence, i.e., an input sample. The values of column y represent the labels. Each of the other columns corresponds to a single amino acid sequence of the historical sequences for a specific target site, where each number represents an amino acid triplet, and the list of 3 numbers in each cell represents the three overlapping 3-grams that are located around the target site.

Table S1. An Example of Generated Sequence Data. Each value represents a triple of amino acids. For a total of 20 amino acids, all combinations of the amino acid triples can be indexed from 1 to 20^3, where the integer is the such index.

y	0	1	2	3	4
0	[602, 1017, 1378]	[602, 1017, 1378]	[602, 1017, 1378]	[602, 1017, 1378]	[2020, 2334, 552]
0	[1078, 312, 903]	[1078, 312, 903]	[1078, 312, 1437]	[1078, 312, 903]	[1078, 312, 903]
0	[461, 477, 2671]	[461, 477, 2671]	[461, 477, 2671]	[461, 477, 2671]	[461, 477, 2671]
1	[1466, 2042, 2173]	[1466, 2042, 2173]	[1466, 2042, 2173]	[1466, 2042, 2173]	[1466, 2042, 2173]
1	[242, 582, 2805]	[242, 582, 2805]	[242, 582, 2805]	[242, 582, 2805]	[242, 582, 2805]
0	[771, 1033, 118]	[771, 1033, 118]	[771, 1033, 118]	[771, 1033, 118]	[771, 1033, 118]
0	[7814, 6853, 2270]	[7814, 6853, 2270]	[7814, 6853, 2270]	[7692, 6853, 2270]	[7814, 6853, 2270]
1	[765, 1837, 2433]	[765, 1837, 2433]	[765, 1837, 2433]	[765, 1837, 2433]	[765, 1837, 2433]
1	[6092, 2395, 472]	[6092, 2395, 472]	[6092, 2395, 472]	[6092, 2395, 472]	[6092, 2395, 472]
0	[7623, 7540, 5624]	[7623, 7540, 5624]	[7623, 7540, 5624]	[7623, 7540, 5624]	[7623, 7540, 5624]
0	[3005, 1114, 752]	[3005, 1114, 752]	[3005, 1114, 752]	[3005, 1114, 752]	[3005, 1114, 1897]
1	[6122, 6331, 5744]	[6122, 6331, 5744]	[6122, 6331, 5744]	[6122, 6331, 5744]	[2462, 4377, 3328]
0	[158, 47, 542]	[158, 47, 542]	[158, 47, 542]	[158, 47, 542]	[158, 47, 542]
0	[3748, 5157, 3064]	[3748, 5157, 3064]	[3748, 5157, 3064]	[3748, 5157, 3064]	[3748, 5157, 3064]
0	[28, 463, 1063]	[28, 463, 1063]	[28, 463, 1063]	[28, 463, 1063]	[28, 463, 1063]
0	[1329, 6663, 7213]	[1329, 6663, 7213]	[2153, 6876, 7213]	[1329, 6663, 7213]	[1329, 6663, 7213]
1	[13, 1259, 2079]	[13, 1259, 2079]	[13, 1259, 2079]	[13, 1259, 2079]	[1048, 2178, 1826]
1	[2270, 2814, 3670]	[2270, 2814, 3670]	[2270, 2814, 3670]	[2270, 2814, 3670]	[2270, 2814, 3670]
1	[3292, 1333, 1064]	[3292, 1333, 1064]	[3292, 1333, 1064]	[3292, 1333, 1064]	[3292, 1333, 1064]
1	[2648, 5437, 7046]	[2648, 5437, 7046]	[2648, 5437, 7046]	[2648, 5437, 7046]	[9047, 9047, 9047]
0	[4051, 1734, 1681]	[4051, 1734, 1681]	[4051, 1734, 1681]	[4051, 1734, 167]	[4051, 1734, 1681]
1	[2129, 880, 870]	[2129, 880, 870]	[2129, 880, 870]	[2129, 880, 870]	[1033, 1166, 132]
1	[1477, 3478, 1800]	[1477, 3478, 1800]	[1477, 3478, 1800]	[1477, 3478, 1800]	[1588, 3913, 3062]
1	[489, 1209, 3761]	[489, 1209, 3761]	[489, 1209, 3761]	[489, 1209, 3761]	[489, 1209, 3761]

3 MUTATION PREDICTION RESULTS OF ALL SITES

The mutation prediction results of all sites are demonstrated in Figure S1. The x-axis denotes the site id, and the y-axis denotes the prediction probabilities of these sites. The detailed information of mutation prediction results of all sites can be found in the online repository https://github.com/ZJUDataIntelligence/TEMPO.

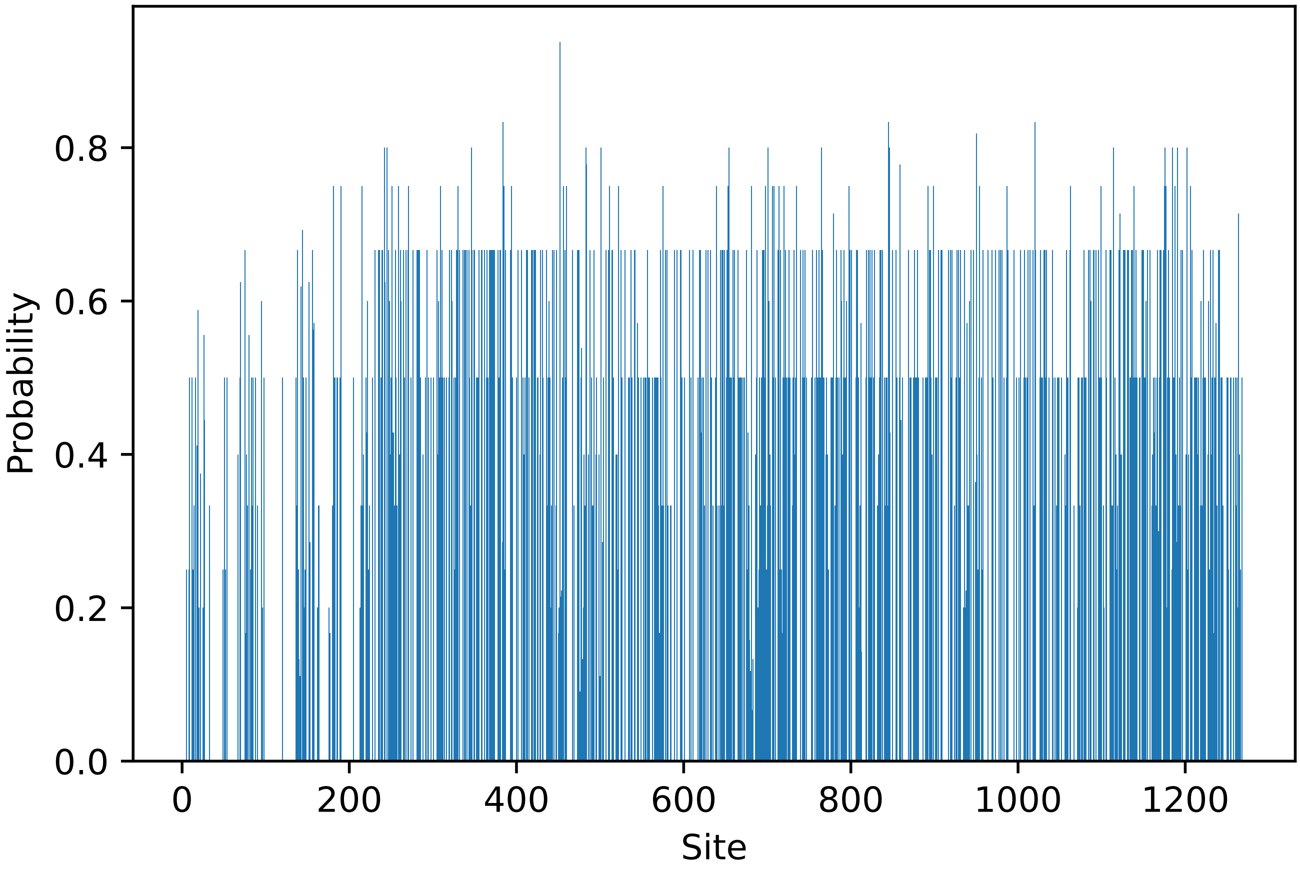

Figure S1: Prediction Probability of All Sites

4 PREDICTION PROBABILITY OF NEW MUTATIONS AFTER FEBRUARY 2022

Table S2 presents the prediction probability of the emergency of new mutations after February 2022 at previously mutated sites. And Table S3 shows the prediction probability of mutations in sites that have never seen mutations.

Table S2. Prediction probability of new mutations after February 2022 at previously mutated sites

Mutation	Prediction probability	# Sampled sequence data
YH145-146-	0.50	12
L212S	-	0
D215E	0.75	8
Y248N	0.60	5
Y248H	0.60	5
G339R	0.67	3
G339S	0.67	3
S371Y	0.67	3
N439	0.60	5
L452M	0.94	16
T547I	-	0

Table S3. Predicted probability of new mutations after February 2022 in sites that have never seen mutations

Mutation	Prediction probability	# Sampled sequence data	Mutation	Prediction probability	# Sampled sequence data
V3G	-	0	F562	-	0
D53	-	0	N641S	-	0
W64R	-	0	A653V	0.75	4
W64L	-	0	N658S	-	0
H69Y	0.50	10	I670V	-	0
K97E	-	0	Q677E	0.43	7
L117	-	0	S735	0.75	4
K147E	0.25	4	T747I	0.50	4
N148T	0.50	2	L858I	0.50	6
F186S	0.50	2	W886L	0.50	4
I210V	-	0	T941S	0.33	6
I233V	-	0	A1020S	0.83	6
G257S	0.33	6	V1129I	-	0
T259A	0.75	4	E1202Q	0.80	5
N354	0.50	2	E1202	0.80	5
R357K	-	0	I1221T	0.50	6
Y449N	-	0	T1231S	0.40	5
Y449H	-	0	C1248	-	0
N460K	0.75	4	S1261Y	0.33	3
F486V	0.40	5			