Methods and Applications

A Novel Matching Pursuit Modeling Strategy Based on Adaptive Fourier Decomposition Theory for Predicting Antigenic Variation of Influenza A (H1N1)

Wei Qu[1,2,&]; Ruihan Chen[3,&]; Yang Wang[1,3,4]; Zhiqi Zeng[5,6,7]; Cheng Gao[8]; Weiqi Pan[1,9]; Tao Qian[10,#]; Chitin Hon[3,9,#]; Zifeng Yang[1,4,#]

ABSTRACT

Introduction: Seasonal influenza poses a significant public health burden, causing substantial morbidity and mortality worldwide each year. In this context, timely and accurate vaccine strain selection is critical to mitigating the impact of influenza outbreaks. This article aims to develop an adaptive, universal, and convenient method for predicting antigenic variation in influenza A(H1N1), thereby providing a scientific basis to enhance the biannual influenza vaccine selection process.

Methods: The study integrates adaptive Fourier decomposition (AFD) theory with multiple techniques — including matching pursuit, the maximum selection principle, and bootstrapping — to investigate the complex nonlinear interactions between amino acid substitutions in hemagglutinin (HA) proteins (the primary antigenic protein of influenza virus) and their impact on antigenic changes.

Results: Through comparative analysis with classical methods such as Lasso, Ridge, and random forest, we demonstrate that the AFD-type method offers superior accuracy and computational efficiency in identifying antigenic change-associated amino acid substitutions, thus eliminating the need for time-consuming and expensive experimental procedures.

Conclusion: In summary, AFD-based methods represent effective mathematical models for predicting antigenic variations based on HA sequences and serological data, functioning as ensemble algorithms with guaranteed convergence.Following the sequence of indicators specified in I, we perform a series of operations on A_1, including feature extension, extraction, and rearrangement, to generate a new input dataset A_1 for the prediction step. With this newly prepared input, we can compute the predicted results as $A_1 W$.

Seasonal influenza remains a significant global public health threat, with the World Health Organization (WHO) estimating 3 to 5 million severe cases and 290,000 to 650,000 deaths annually (*1*). The predominant circulating strains — influenza A virus subtype H1N1 [A(H1N1)], A(H3N2), and B(Victoria) — undergo antigenic drift due to amino acid substitutions in the hemagglutinin (HA) protein. These molecular changes enable the virus to evade host immunity, resulting in seasonal outbreaks (*2*–*3*). Traditional serologic assays, such as hemagglutination inhibition (HI), are employed to monitor antigenic changes but are labor-intensive, costly, and require live virus isolation (*4*). Consequently, a sequence-based strategy to predict antigenic variants would represent a more efficient alternative (*5*).

Several machine learning models have been developed for HA sequence-based antigenicity prediction, including support vector machines (SVM), multi-task learning sparse group lasso (MTL-SGL), iterative filtering models, and ridge regression. These approaches demonstrate robust performance in high-dimensional data classification, integrating multiple features with numerical weighting (*6*–*8*). However, these models exhibit limitations in handling dynamic data and nonlinear relationships, rendering predictions susceptible to noise, missing values, and feature correlation.

In this article, we introduce a matching pursuit model based on adaptive Fourier decomposition (AFD) theory for predicting influenza antigenic variation, using H1N1 as an exemplar. Inspired by (*9*) and (*10*), our model offers three distinct advantages: Adaptivity and efficiency via an AFD maximum selection that mitigates overfitting on small datasets; Nonlinearity and interpretability through capturing epistatic effects between amino acid changes and spatial positions; Robustness via feature screening,

bootstrapping, and orthogonal projection for dual-site interactions.

METHODS

Matching Pursuit Model Based on Adaptive Fourier Decomposition Theory

This section develops a quantitative model to predict antigenic distances from HA protein sequences. We denote A as the independent features and Y as the target variable. Details on the matching pursuit model and prediction procedure are provided in the Supplementary Material (available at https://weekly.chinacdc.cn/).

In this section, we outline the specific steps of the model algorithm, which are divided into two main phases: training and predicting, which are shown in Table 1 and Table 2, respectively.

Assuming the execution of the above algorithm stops at step $j = p_\epsilon (\leq p)$, and we obtain the parameter set $X = (x_1, \cdots, a_{I_{p_\epsilon}})$ for the training model and the index set $I = (I_1, \cdots, I_{p_\epsilon})$. Let $B = (b_1, \cdots, b_{p_\epsilon})$ represent the orthonormal matrix, and $A = (a_{I_1}, \cdots, a_{I_{p_\epsilon}})$ represent the rearranged matrix of A according to I. We can compute $W_{p_\epsilon \times p_\epsilon}$ using $B = A\,W$, which gives us the parameter set $W = W\,X^t$ for prediction model. The subsequent algorithm will help us derive the parameter set for the prediction model and present the prediction results.

Both algorithms generate sequence data through feature expansion, which can lead to a high-dimensional space and increased overfitting risk — especially when higher-order terms are included. However, our model mitigates this via a maximum selection principle and by applying expansion to both training and testing sets. To balance enhanced

TABLE 1. Matching pursuit algorithm — training model.

Step	Process
Input	sequence data $A_{q\times p}=(a_1,...,a_q)$ and antigenic data $Y_{q\times 1}$
Output	the parameter set X, the index set I and the result $\tilde{Y}_{q\times 1}$
0	Initialize ε>0, j=1
	$b_k \leftarrow a_k/\|\|a_k\|\|$, k = 1,⋯,p
	$I_1 \leftarrow \text{argmax}_k \|<Y,b_k>\|^2$
	$b_1 \leftarrow a_{I1} / \|\|a_{I1}\|\|$
	$x_1 \leftarrow <Y,b_1>$
	$\tilde{Y} \leftarrow <Y,b_1>b_1$
	$\text{energy} \leftarrow \|x_1\|^2$
1	While energy≥ε && j<p do
2	$j \leftarrow j + 1$
3	$b_k \leftarrow Q_{bj-1}(b_k)/\|\|Q_{bj-1}(b_k)\|\|$, k = 1,⋯,p
4	$I_j \leftarrow \text{argmax}_k \|<Y,b_k>\|^2$
5	$b_j \leftarrow b_{Ij}$
6	$x_j \leftarrow <Y,b_j>$
7	$\tilde{Y} \leftarrow \tilde{Y} + <Y,b_j>b_j$
8	$\text{energy} \leftarrow \|x_j\|^2$
9	End while

TABLE 2. Matching pursuit algorithm — predicting model.

Step	Process
Input	X, I, W, and new sequence data, denoted by $A_{q1\times p}$
Output	prediction result, denoted by $\tilde{Y}_{q1\times 1}$
0	extract and rearrange a subset of $A_{q1\times p}$ according to I; then obtain $\tilde{A}_1$with size $q_1\times p_\varepsilon$
1	compute $W = W\,A^t$
2	compute $\tilde{Y}_{q1\times 1} = \tilde{A}\,W$

prediction accuracy with the increased computational cost of higher dimensions, we randomly select a small subset of features, choose an appropriate expansion degree (e.g., 2nd or 3rd), and then perform random feature sampling with replacement. The final prediction is obtained by averaging across all iterations, leveraging ensemble methods similar to those used in random forests.

Validation Examples

The dataset description is provided in the Supplementary Materials. In this section, we first present the model's training and prediction results, followed by an evaluation using multiple performance metrics. We then discuss the reliability of key sites identified by the model, particularly in the context of antigenic variation. We employ two primary evaluation metrics to assess model effectiveness: root mean square error (RMSE) and F1-score, defined as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{Y}_i)^2}$$

where Y represents the true value and $\tilde{y}$ represents the predicted result

For each analytical task, we employ Algorithm 1 for training and Algorithm 2 for prediction. We benchmark our approach against five classical methods: Random Forest (RF), Support Vector Regression (SVR), Lasso, Gradient Boosting (GB), and Elastic Net (EN). Our proposed model is Matching Pursuit Method (MP).

RESULTS

Model Evaluation

We established epsilon values of 0.1, 0.01, 0.01, 0.001, and 0.01, with bootstrap samples of 30, 5, 5, 2, and 15 across the five tasks, respectively. Each task incorporated 70, 80, 70, 80, and 80 observations drawn with replacement from the original dataset. Subsequently, we calculated the mean for each of these samples. From a theoretical perspective, as the number of selected observations decreases, the number of bootstrap samples should increase proportionally. The evaluation metrics for the training model are presented in Table 3.

The five tasks above demonstrate that our method performs robustly across these datasets. The approach proves effective both in capturing positive events, such as site variations, and in optimizing the balance between accuracy and recall rate.

Figure 1 displays the MP model's training results for antigenic distance prediction, where blue dots closer to the red line indicate superior performance. We subsequently applied Kernel Density Estimation (KDE) with a bandwidth of 0.5 to generate smooth density curves for both predicted and actual data. The substantial overlap between these curves reveals similar distributions and minimal bias. As illustrated in Figure 2, this alignment across datasets confirms the model's strong generalization capabilities, consistency, and robustness.

The evaluation metrics for the prediction model are presented in Table 4.

The prediction results across the five tasks above reveal that, while our model demonstrates strong performance during training, the prediction outcomes still present opportunities for improvement. Despite systematic efforts to optimize parameters and refine the input dataset during model development, certain aspects remain suboptimal. Nevertheless, these numerical results provide valuable reference points for subsequent research endeavors.

TABLE 3. Comparison of training performance between classical models and AFD-based predictive methods on five H1N1 prediction tasks.

Methods	Task 1		Task 2		Task 3		Task 4		Task 5	
	RMSE	F1-score	RMSE	F1-score	RMSE	F1-score	RMSE	F1-score	RMSE	F1-score
RF	0.624	0.730	0.380	0.899	0.453	0.909	0.326	0.984	0.366	0.816
SVR	0.203	0.955	0.343	0.956	0.506	0.890	0.323	0.968	0.335	0.883
Lasso	1.317	0.543	1.322	0.867	1.635	0.113	0.905	0.878	1.340	0.520
GBR	0.763	0.730	0.708	0.867	0.790	0.808	0.561	0.878	0.433	0.768
ENG	0.519	0.909	0.597	0.932	0.627	0.863	0.371	0.984	0.341	0.816
MP	**0.149**	**0.978**	**0.296**	**0.963**	**0.312**	**0.939**	**0.195**	**1.000**	**0.261**	**0.930**

Note: The bolded values highlight the best performance scores across different models for each H1N1 prediction task.
Abbreviation: RF=random forest; SVR=support vector regression; GBR=gradient boosting regression; ENG=elastic net; MP=matching pursuit method; RMSE=root mean square error.

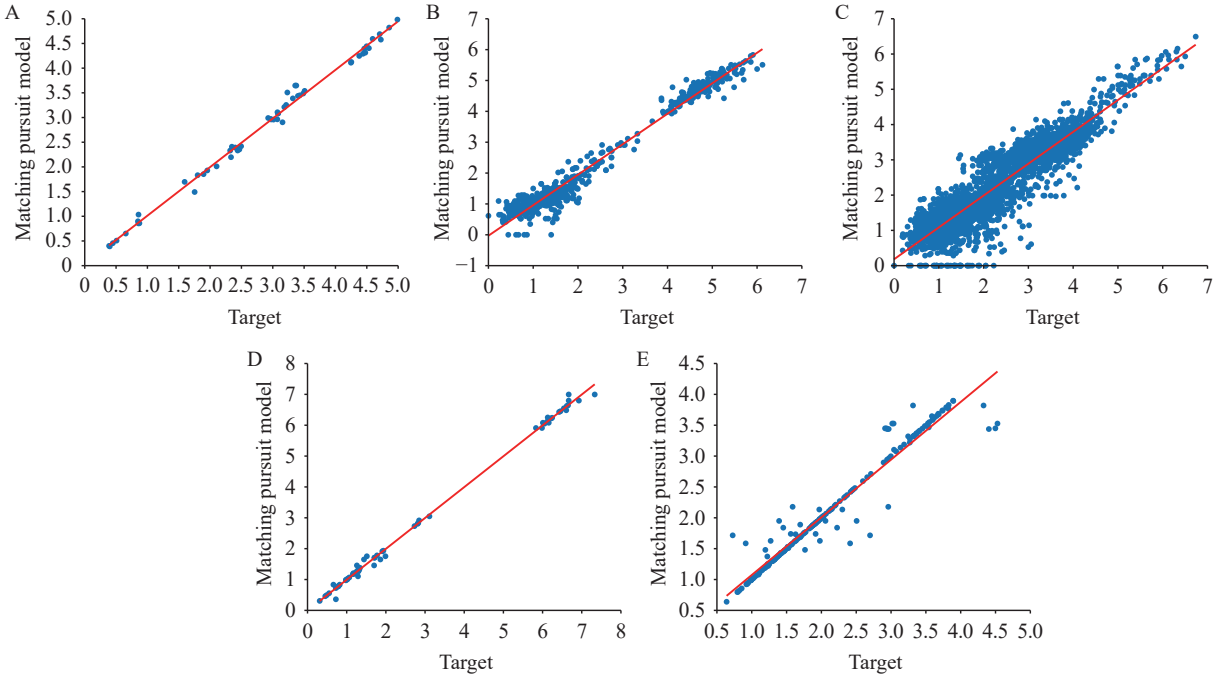


FIGURE 1. Training results of the MP model for antigenic distance prediction across (A–E) Tasks 1–5.
Note: The X-axis represents the ground truth antigenic distance, and the Y-axis shows the predicted values. The red diagonal line is the correlation line.
Abbreviation: MP=matching pursuit method.

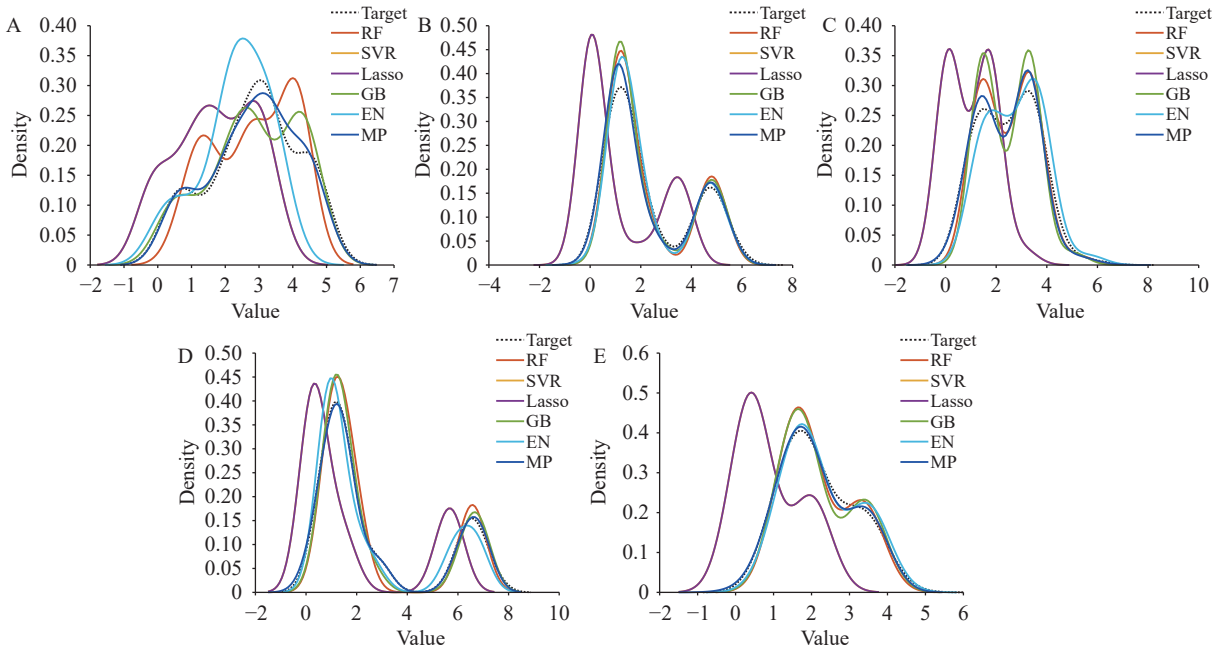


FIGURE 2. Training results of the classical and MP model represented through Kernel Density Estimation (KDE) distributions of predicted and actual antigenic distance values across (A–E) Task 1–5.
Note: The X-axis denotes the antigenic distance, and the Y-axis indicates the density. Each line corresponds to a different model.
Abbreviation: MP=matching pursuit method.

Figure 3 illustrates the prediction results for antigenic distance using the MP model. The proximity of blue dots to the red line indicates prediction accuracy. Figure 4 displays the KDE results for all six methods, demonstrating that our approach yields superior testing outcomes. The degree of overlap with

the target curve directly corresponds to prediction performance quality.

Analysis on Amino Acid Site

In this section, we conducted a systematic screening and evaluation of critical amino acid sites within the model. The top 50 amino acid sites with the highest contribution were selected for model fitting in each task. Task 1 comprised 8 single sites and 34 coupled sites, task 2 included 13 single sites and 37 coupled sites, task 3 contained 12 single sites and 38 coupled sites, task 4 had 8 single sites and 32 coupled sites, and task 5 consisted of 7 single sites and 43 coupled sites. Notably, coupled sites consistently represented a higher proportion in feature selection across all tasks, ranging from 74–86 percent (Table 5 and Table 6).

We identified 21, 29, 39, 37, and 53 amino acid mutations in tasks 1–5, with 16, 20, 29, 22, and 28 sites respectively associated with antigenic epitopes (Table 7 and Figure 5). These findings suggest that mutations at these positions may significantly alter antigenicity and contribute to antigenic drift. Notably, certain amino acid positions appeared repeatedly in coupled-site mutations, such as positions 216 and 186

TABLE 4. Comparison of predicting performance between classical models and AFD-based predictive methods on five H1N1 prediction tasks.

Methods	Task 1		Task 2		Task 3		Task 4		Task 5	
	RMSE	F1-score	RMSE	F1-score	RMSE	F1-score	RMSE	F1-score	RMSE	F1-score
RF	0.678	**0.942**	0.573	0.891	0.523	0.905	0.405	**0.941**	0.556	0.817
SVR	1.065	0.821	0.757	0.913	0.570	0.889	0.799	0.898	0.526	0.871
Lasso	1.315	0.517	1.301	0.891	1.617	0.111	1.334	0.806	1.414	0.164
GBR	0.942	0.826	0.747	0.891	0.786	0.827	1.582	0.570	0.661	0.796
ENG	0.653	0.921	0.780	0.927	0.610	0.877	0.456	0.962	0.546	0.844
MP	**0.582**	**0.942**	**0.478**	**0.944**	**0.513**	**0.914**	**0.403**	**0.941**	**0.416**	**0.915**

Note: The bolded values highlight the best performance scores across different models for each H1N1 prediction task.
Abbreviation: RF=random forest; SVR=support vector regression; GBR=gradient boosting regression; ENG=elastic net; MP=matching pursuit method; RMSE=root mean square error.

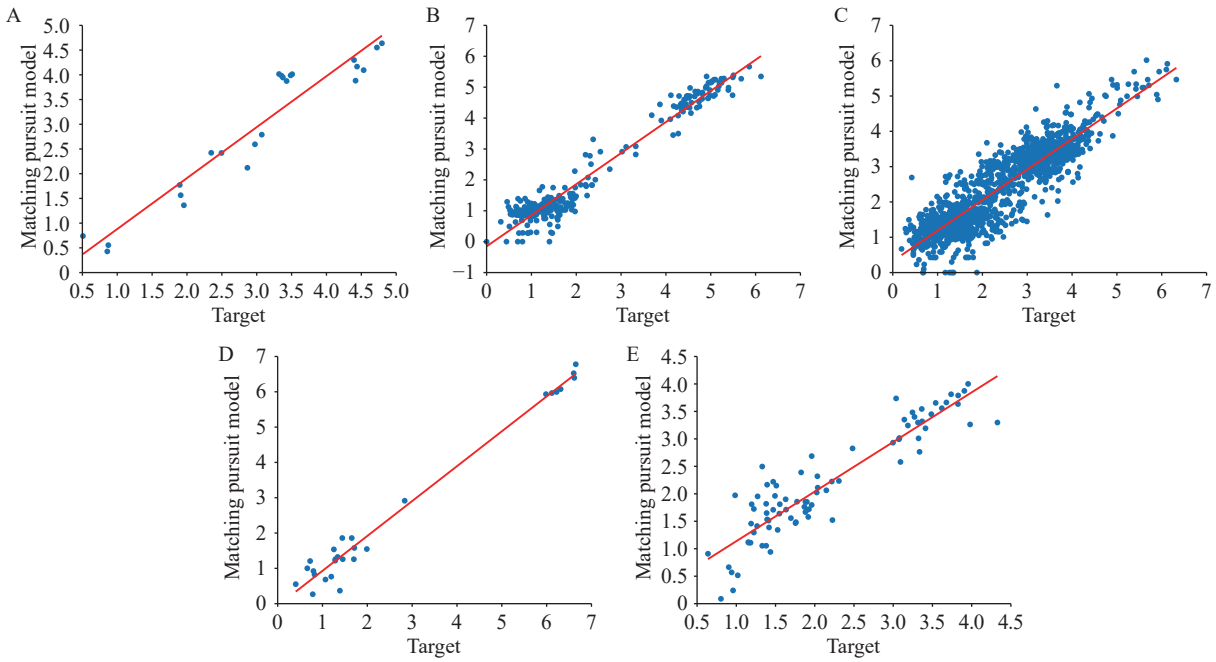


FIGURE 3. Predicting results of the MP model for antigenic distance prediction across (A–E) Task 1–5.
Note: The X-axis represents the ground truth antigenic distance, and the Y-axis shows the predicted values. The red diagonal line is the correlation line.
Abbreviation: MP=matching pursuit method.

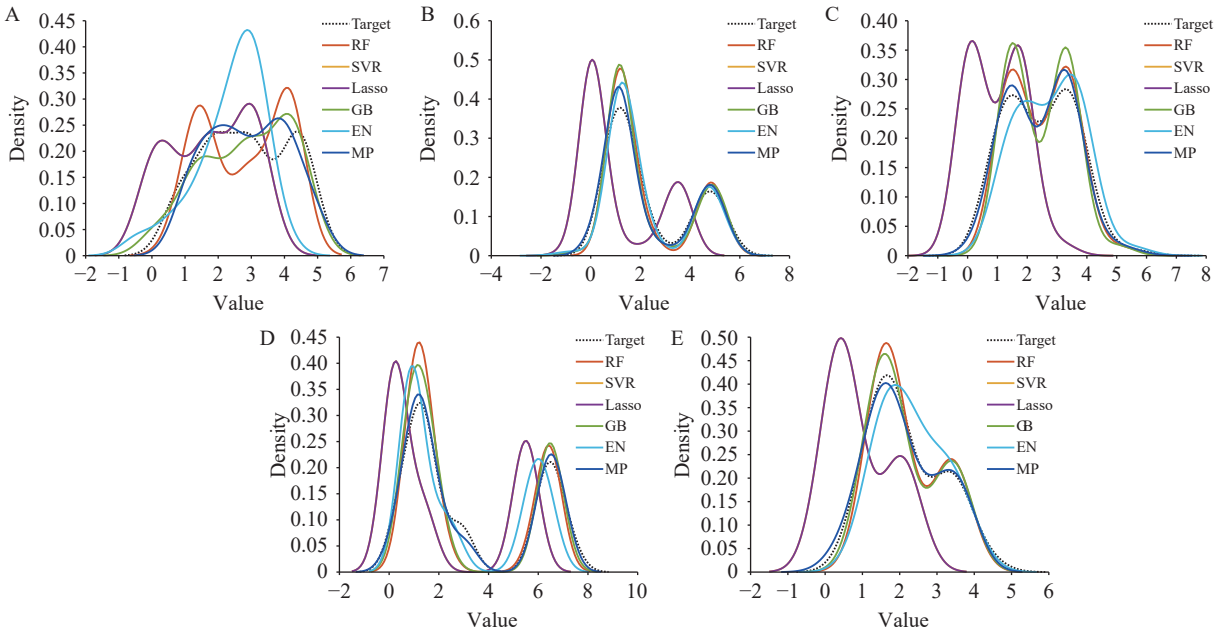


FIGURE 4. Predicting results of the classical and MP model represented through KDE distributions of predicted and actual antigenic distance values across (A–E) Task 1–5.
Note: The X-axis denotes the antigenic distance, and the Y-axis indicates the density. Each line corresponds to a different model.
Abbreviation: KDE=kernel density estimation; MP=matching pursuit method.

TABLE 5. Top single amino acid sites identified for their high contribution to antigenic changes within each task based on the MP model (Single Site).

Task 1 (8)	Task 2 (13)	Task 3 (12)	Task 4 (8)	Task 5 (7)
54	43	43	51	9
56	66	57	120	34
71	74	82	155	49
121	84	132	186	77
128	89	141	211	81
135	125	186	216	93
186	141	187	260	95
187	153	189	272	
	163	190		
	187	222		
	215	252		
	222	315		
	253			

Note: The number after Task No. is the important feature number.
Abbreviation: MP=matching pursuit method.

in task 1, 253 in task 2, 187 and 141 in task 3, 211 in task 4, and 209 and 35 in task 5. The recurrence of these mutations in both single-site and coupled-site analyses indicates their substantial impact on antigenic properties (Table 6 and Figure 6).

Based on the results shown in Table 6 and Figure 5, we have identified both commonalities and differences across individual tasks. Certain amino acid sites consistently appear in multiple tasks, such as the 153 site in the Sa region, which is identified as critical in almost all tasks, suggesting its central role in antigenic variation. Conversely, some loci appear exclusively in specific tasks, reflecting the diversity of antigenic variations that may be influenced by different datasets or model conditions.

Finally, we summarized and deduplicated the amino acids in six antigenic epitopes (Ca, Cb, Pa, Pb, Sa, and Sb) selected from the five tasks. A total of 12 residues are present in the Ca antigenic epitope, 13 in the Cb antigenic epitope, 8 in the Pa antigenic epitope, 4 in the Pb antigenic epitope, 11 in the Sa antigenic epitope, and 12 in the Sb antigenic epitope. All residues were visualized on both trimeric and monomeric structures of the influenza HA protein (PDB: 3UBE) using PyMOL (Figure 7).

The identification of these key sites provides valuable insights for elucidating antigenic variation mechanisms and serves as a critical reference for vaccine design. Specifically, optimizing vaccine formulations to target these frequently occurring critical sites could substantially enhance vaccine efficacy against emerging viral strains.

TABLE 6. Top coupled amino acid sites identified for their high contribution to antigenic changes within each task based on the MP model.

Task No.	Two Site									
Task 1 (34)	187–222 56–193 141–157 135–160	135–186 54–56 135–141 160–216	121–216 56–216 186–253 157–272	56–253 36–186 128–253 135–222	186–216 153–160 71–135	71–130 128–186 74–135	71–186 128–193 160–324	193–216 193–253 36–157	54–272 74–141 36–216	121–187 36–193 56–130
Task 2 (37)	69–125 2–315 89–153 125–253	187–253 84–187 273–324 3–82	153–187 252–253 2–163 43–187	43–125 74–222 2–72 43–73	153–253 43–183 2–84 69–190	187–215 69–175 84–253 2–43	222–273 153–209 166–253 43–253	74–141 72–315 153–163	125–183 163–187 175–253	3–253 208–253 66–215
Task 3 (38)	187–189 186–187 170–194 35–194	183–253 82–187 141–193 35–73	69–269 267–273 160–193 146–187	186–189 194–209 120–141 267–315	73–128 141–194 141 187–252	189–271 183–186 132–153 166–209	267–290 82–190 68–141 187–215	132–141 187–190 73–189 187–315	74–183 194–208 112–209	74–189 84–141 73–82
Task 4 (32)	71–162 17–260 72–134 129–222	45–211 162–260 84–215 94–1	120–272 84–228 3–228	56–112 155–228 32–47	38–47 271–283 43–72	47–71 168–170 211–260	38–211 211–250 72–250	47–250 17–47 32–276	211–298 94–129 161–271	32–43 38–250 61–168
Task 5 (43)	43–130 74–156 127–239 83–262 96–127	35–186 138–183 19–187 197–227 209–298	36–130 120–128 61–178 3–197 183–190	89–129 83–109 85–161 36–209	207–260 43–129 19–69 161–19	109–209 71–129 35–205 89–239	129–166 179–239 179–209 73–178	36–129 71–179 51–179 166–179	35–178 183–187 128–197 128–186	38–45 84–262 191–274 35–170

Note: The number after Task No. is the important feature number.
Abbreviation: MP=matching pursuit method.

TABLE 7. Antigenic sites and corresponding amino acid positions within the HA1 epitope identified as critical for antigenic changes across tasks based on the MP model.

Antigenic sites	Task 1-aa	Task 2-aa	Task 3-aa	Task 4-aa	Task 5-aa
Sa	121, 153, 157,160	125, 153, 163	120, 153, 160	120, 155, 161, 162	120, 156, 161
Sb	186, 187, 193	187, 190, 208, 209	186, 187, 189, 190, 193, 194, 208, 209	186, 211	186, 187, 190, 191, 197, 207, 209
Ca	141, 216, 222	141, 166, 215, 222	141, 146, 166, 170, 215, 222	142, 168, 170, 215, 216, 222	138, 166, 170, 205, 239
Cb	54, 71, 74, 253	72, 73, 74, 82, 84, 89, 253	68, 73, 74, 82, 84, 253	71, 72, 84, 260	71, 73, 74, 84, 85, 89, 260, 262
Pa	272	43, 273	43, 269, 271, 273	43, 271, 276, 283	43, 274
Pb	36		35, 290	38	35, 36, 38

Abbreviation: MP=matching pursuit method; HA=hemagglutinin.

DISCUSSION

This article introduces a novel approach for predicting antigenic variations of H1N1 influenza A — the MP model. Traditionally, antigenic variation prediction relies on extracting protein sequences and serological data, followed by applying regression-based models to infer the antigenic characteristics of novel viral protein sequences. In contrast, this study incorporates AFD theory as a key component, offering an alternative analytical perspective that aims to enhance predictive performance and interpretability.

The proposed method demonstrates several significant advantages. First, the algorithm leverages AFD to dynamically select optimal basis functions, which enhances its capacity to capture nonlinear relationships in antigenic data. This flexibility effectively mitigates issues such as overfitting, a common challenge in high-dimensional datasets with sparse labels. Second, compared with traditional regression techniques, the model offers improved interpretability, superior computational efficiency, and reduced complexity, making it particularly suitable for large datasets and real-time applications. Furthermore, the model's applicability extends beyond H1N1 influenza A, with preliminary results suggesting its utility for other influenza subtypes such as H3N2 and Influenza B, and its potential adaptability to other viral families. Notably, this study also incorporates dual-site synergy considerations, identifying key site interactions from five publicly available datasets.

Empirical evaluations on these datasets indicate that the model performs well across various metrics, often outperforming baseline methods. However, deeper

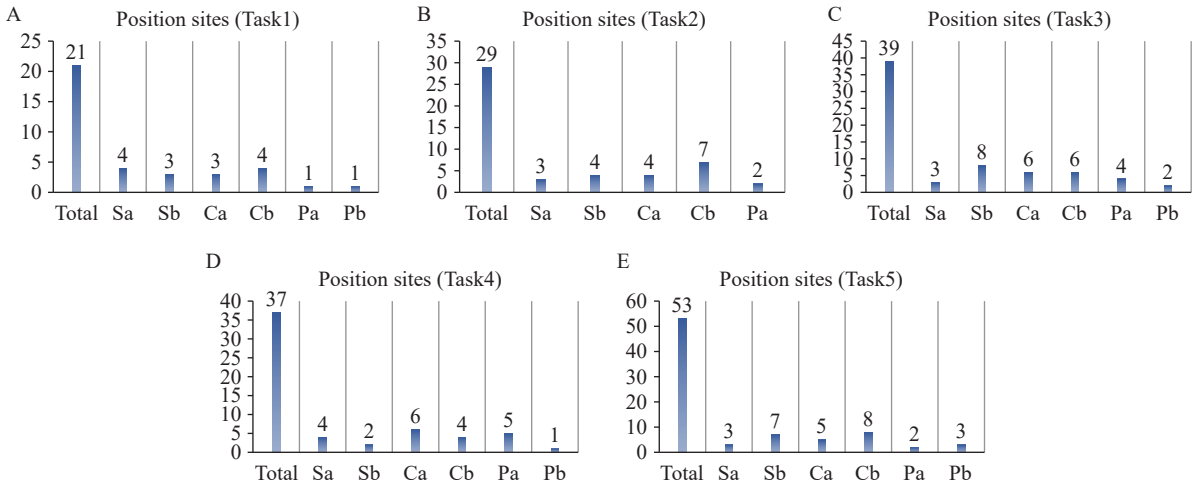


FIGURE 5. Bar charts illustrating the distribution of identified amino acid mutations across antigenic sites (Sa, Sb, Ca, Cb, Pa, and Pb) for (A–E) Tasks 1–5.

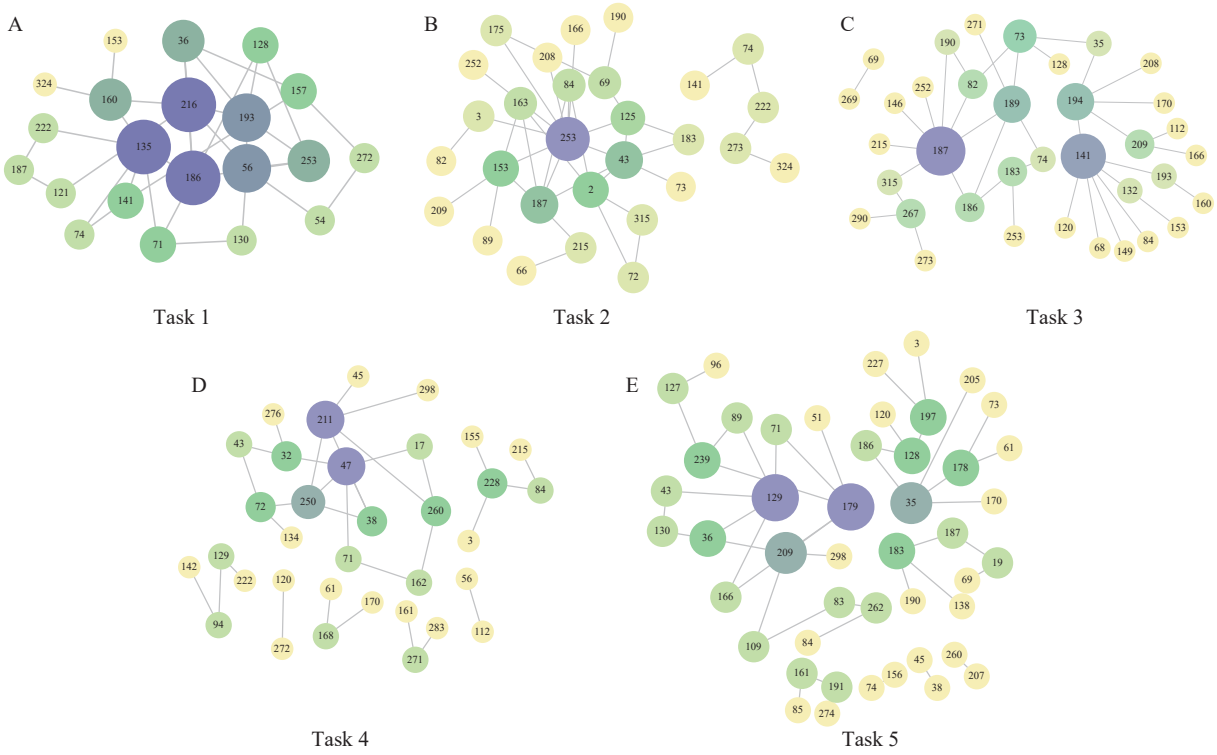


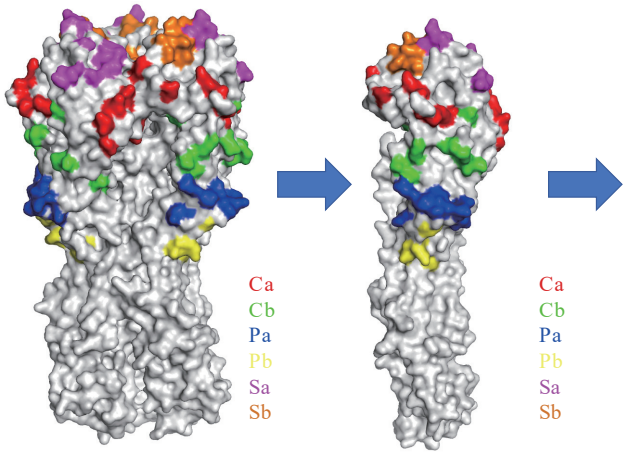
FIGURE 6. Network diagram of two-site interactions for (A–E) Task 1–5.

analysis has revealed certain areas requiring improvement. For example, while the algorithm exhibits strengths in computational efficiency and generalization, its sensitivity to capturing subtle antigenic shifts could be further refined.

Future efforts will focus on integrating advanced feature engineering to capture domain-specific viral protein properties and exploring ensemble learning to enhance predictive robustness. We also plan to collaborate with virology experts on cell-based experiments to validate our predictions and support applications in vaccine design and epidemiological forecasting. This comprehensive approach aims to refine our methodology and contribute to addressing complex challenges in influenza and broader virology research.

Conflicts of interest: No conflicts of interest.

Funding: Supported by Major Project of Guangzhou National Laboratory, (Grant No. GZNL2024A01004), the National Natural Science



Epitope	Number	Site
Ca	12	138, 141, 142, 146, 166, 168, 170, 205, 215, 216, 222, 239
Cb	13	54, 68, 71, 72, 73, 74, 82, 84, 85, 89, 253, 260, 262
Pa	8	43, 269, 271, 272, 273, 274, 276, 283
Pb	4	35, 36, 38, 290
Sa	11	120, 121, 125, 153, 155, 156, 157, 160, 161, 162, 163
Sb	12	186, 187

FIGURE 7. The selected amino acids of six antigenic sites (i.e., Ca, Cb, Pa, Pb, Sa, and Sb) of H1 (A/California/04/2009; PDB 3UBE).

Foundation of China (Grant No. 82361168672), the Science and Technology Development Fund of Macau SAR (Grant No. FDCT 0111/2023/AFJ, 0155/2024/RIA2, 005/2022/ALC, 0128/2022/A, 0020/2023/RIB1), National Key Research and Development Program of China (Grant No. 2024YFE0214800), Self-supporting Program of Guangzhou Laboratory (Grant No. SRPG22-007), National Key Research and Development Program of China (Grant No. SQ2024YFE0202244), Engineering Technology Research (Development) Center of Ordinary Colleges and Universities in Guangdong Province (Grant No. 2024GCZX010).

doi: 10.46234/ccdcw2025.078

[#] Corresponding authors: Tao Qian, tqian@must.edu.mo; Chitin Hon, cthon@must.edu.mo; Zifeng Yang, yang_zifeng@gzlab.ac.cn.

[1] State Key Laboratory of Respiratory Disease, National Clinical Research Center for Respiratory Disease, Guangzhou Institute of Respiratory Health, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou City, Guangdong Province, China; [2] College of Sciences, China Jiliang University, Hangzhou City, Zhejiang Province, China; [3] Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, Macao Special Administrative Region, China; [4] Guangzhou National Laboratory, Guangzhou City, Guangdong Province, China; [5] Guangzhou key laboratory for clinical rapid diagnosis and early warning of infectious diseases, KingMed School of Laboratory Medicine, Guangzhou Medical University, Guangzhou City, Guangdong Province, China; [6] Engineering Technology Research Center of Intelligent Diagnosis for Infectious Diseases in Guangdong Province, Guangzhou City, Guangdong Province, China; [7] Guangdong Provincial Engineering Research Center for Early Warning and Diagnosis of Respiratory Infectious Diseases, Guangzhou City, Guangdong Province, China; [8] Department of Electrical Engineering & Computer Science, College of Engineering, University of Missouri, Columbia, MO, USA; [9] Respiratory Disease AI Laboratory on Epidemic and Medical Big Data Instrument Applications, Faculty of Innovation Engineering, Macau University of Science and Technology, Macao Special Administrative Region, China; [10] Macau Center for Mathematical Sciences, Macao University of Science and Technology, Macao Special Administrative Region, China.
[&] Joint first authors.

Submitted: December 12, 2024
Accepted: March 28, 2025
Issued: April 04, 2025

REFERENCES

1. World Health Organization. Influenza (seasonal). 2024. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). [2024-8-30].
2. Krammer F, Smith GJD, Fouchier RAM, Peiris M, Kedzierska K, Doherty PC, et al. Influenza. Nat Rev Dis Primers 2018;4(1):3. https://doi.org/10.1038/s41572-018-0002-y.
3. Carrat F, Flahault A. Influenza vaccine: the challenge of antigenic drift. Vaccine 2007;25(39-40):6852 – 62. https://doi.org/10.1016/j.vaccine.2007.07.027.
4. CDC. CDC's World Health Organization (WHO) collaborating center for surveillance, epidemiology and control of influenza. 2024. https://www.cdc.gov/flu/php/who-collaboration/index.html. [2024-8-6].
5. Houser K, Subbarao K. Influenza vaccines: challenges and solutions. Cell Host Microbe 2015;17(3):295 – 300. https://doi.org/10.1016/j.chom.2015.02.012.
6. Liao YC, Lee MS, Ko CY, Hsiung CA. Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. Bioinformatics 2008;24(4):505 – 12. https://doi.org/10.1093/bioinformatics/btm638.
7. Li L, Chang D, Han L, Zhang XJ, Zaia J, Wan XF. Multi-task learning sparse group lasso: a method for quantifying antigenicity of influenza A (H1N1) virus using mutations and variations in glycosylation of Hemagglutinin. BMC Bioinformatics 2020;21(1):182. https://doi.org/10.1186/s12859-020-3527-5.
8. Sun HL, Yang JL, Zhang T, Long LP, Jia K, Yang GH, et al. Using sequence data to infer the antigenicity of influenza virus. mBio 2013;4(4):e00230 – 13. https://doi.org/10.1128/mBio.00230-13.
9. Qu W, Hon CT, Zhang YQ, Qian T. Matrix pre-orthogonal matching pursuit and pseudo-inverse. arXiv preprint arXiv:2412.05878, 2025.
10. Hon C, Liu ZG, Qian T, Qu W, Zhao JM. Trends by adaptive Fourier decomposition and application in prediction. Int J Wavelets, Multiresolut Inf Process 2024;22(5):2450014. https://doi.org/10.1142/S0219691324500140.

SUPPLEMENTARY MATERIAL

Data Description

This study utilized serologic data for H1N1 viruses, comprising 2,030 HI titers generated from 153 viruses and 97 serum samples, along with 13,591 non-identical amino acid sequences of HA proteins, accessible at https://github.com/InfluenzaSystemsBiology/MTL-SGL. Analysis of swine-origin influenza viruses (SOIVs) collected from humans between 1990 and 2010 revealed that their HA and neuraminidase (NA) genes belong to the triple-reassortant swine-origin influenza virus (tr-SOIV) lineage, which evolved from classical swine-origin influenza virus (cSOIV) A(H1N1). The tr-SOIV HA genes form two distinct clusters: H1gamma (predominantly east of the Mississippi River) and H1beta (west of the Mississippi River). Seasonal human H1N1 viruses (1977–2009) were characterized using HI assays, involving 115 virus isolates, 77 serum isolates, and 1,882 measurements to correlate antigenic dynamics with molecular evolution. Additionally, swine H1N1 viruses from 2008 were characterized through genome sequencing and serological cross-reactivity analysis to elucidate genetic diversity and antigenic properties.

Our analysis incorporated five datasets, denoted as (A_1, Y_1), (A_2, Y_2), (A_3, Y_3), (A_4, Y_4), (A_5, Y_5). Matrices A_m represent HA protein sequences, while vectors Y_m represent antigenic distances, where $m = 1, \cdots 5$. The dimensions of A_m are 78 × 167, 861 × 167, 4,950 × 167, 91 × 167, and 276 × 167, respectively. The dimensions of vectors Y_m correspond to the number of rows in matrix A_m. For each dataset, amino acid substitutions and antigenic distances for virus pairs were determined using established protocols. Specifically, amino acid substitutions were quantified using a binary coding schema for all HA sequence pairs. Antigenic distances for corresponding virus pairs were calculated based on their HI titers against different antisera using a low-rank matrix completion method. Each dataset was randomly partitioned into two segments: 70 percent allocated for model training to establish underlying patterns and relationships, and the remaining 30 percent reserved as a testing set to evaluate model performance on unseen data.

Formulation of Matching Pursuit Model

Suppose $A = (a_1, a_2, \ldots, a_p)$, where $a_l \in R^q$, $l = 1, 2, \cdots p$, are the non-zero column vectors of A and R^q is the Euclidean space with dimension q. Let $Y \in R^q$. Leveraging the idea of polynomial regression, we perform feature expansion on A. Through feature expansion, the elements of A can be transformed into higher-order features such as squared terms, cubic terms, or even higher-order terms. Additionally, this process includes interaction terms between the elements of A, known as feature interactions. This enhancement significantly boosts the model's expressive power, enabling us to effectively capture complex nonlinear relationships when addressing real-world problems. Without loss of generality, we still use the notation $A = (a_1, a_2, \ldots, a_p)$ to represent the result after completing the feature extension.

The objective is to identify a re-ordered subset of A, denoted as $\mathrm{A} = (a_{I_1}, \cdots, a_{I_{p'}}), p' \leq p$, and a row vector $\mathrm{X} = (\mathrm{x}_1, \cdots, \mathrm{x}_{\mathrm{p1}})$ such that

$$||Y - \mathrm{AX}^{\mathrm{t}}||$$

is minimized, where $||\cdot||$ is the Euclidean norm and X^{t} is the transpose of X.

Let $B = (b_1, b_2, \ldots, b_p)$ be the normalization of vectors A, i.e., $b_l = a_l / ||a_l||_2$, $l = 1, 2, \cdots, p$. Denote by Q_{bl} the Gram-Schmidt orthogonalization with respect to any vector bl. That is

$$Q_{bl}(a_l) = a_l - \langle a_l, b_l \rangle b_l.$$

For a set of p orthogonal vectors $\mathrm{B} = \mathrm{b}_1, \cdots, \mathrm{b}_q$, the following relationship holds:

$$Q_{b_p}\left(Q_{b_{p-1}}\left(\cdots\left(Q_{b_1}(a_l)\right)\cdots\right)\right) \triangleq Q_{b_p} \circ Q_{b_{p-1}} \circ \cdots \circ Q_{b_1}(a_l)$$

Corresponding to the description in Section 2, we consider A as the kernel functions, B as the normalized kernel functions, and B as the Takenaka-Malmquist system. Through this framework, we can apply the discrete form of the maximal selection principle (MSP).

For the initial case where $l = 1$, we select I_1 and a_{I1} according to:

$$\arg max\{|\langle Y, \frac{a_k}{\| a_k \|}\rangle|^2 : k = 1, \cdots, p\}$$

We then set $b_1 = \frac{a_{I_1}}{\|a_{I_1}\|}$ and $b_k^l = b_k^1 = Q_{b_1}\left(a_k^{l-1}\right) = Q_{b_1}\left(a_k^0\right), k = 1, \cdots, p,$, where we adopt the notation $a_k^0 = a_k$. This yields the first parameter $x_1 = \langle Y, b_1\rangle$ for our model, where $\langle ., .\rangle$ represents the Euclidean inner product.

Next, we discuss $l = 2$. Select I_2 and a_{I2} according to

$$\arg max\{|\langle Y, \frac{b_k^1}{\| b_k^1 \|}\rangle|^2 : k = 1, \cdots, p\}$$

Naturally, $I_2 \neq I_1$. If the maximum value is zero, the selection process terminates at this step. Otherwise, set $b_1 = \frac{b_{I_2}^l}{\|b_{I_2}^l\|}$ and $b_k^l = b_k^2 = Q_{b_2}\left(b_k^1\right), k = 1, \cdots, p$. Furthermore, we derive the second parameter $x_2 = \langle Y, b_2\rangle$ for our model.

Inductively, for any integer $l > 2$, we ultimately obtain an orthonormal system $b_1, , b_{p'}$ and an integer $p' \leq p$, where p' can be computed by

$$\arg max\{|\langle Y, \frac{b_k^{p'}}{\| b_k^{p'} \|}\rangle|^2 < \epsilon : k = 1, \cdots, p\}$$

for a predetermined ε. Controlling the sample size (i.e., the size of q), along with feature expansion, ensures that the integer $p'(\leq p)$ is attained. Denote by $X = (x_1, \cdots, x_{p'})$, where $x_1 = \langle Y, b_l\rangle, l = 1, , p'$. Additionally, we have

$$b_l = \frac{Q_{b_{l-1}} \circ \cdots \circ Q_{b_1}\left(a_{I_1}^0\right)}{\| Q_{b_{l-1}} \circ \cdots \circ Q_{b_1}\left(a_{I_1}^0\right) \|} = \frac{Q_{b_{l_1}}\left(b_{I_l}^{l-2}\right)}{\| Q_{b_{l-1}}\left(b_{I_l}^{l-2}\right) \|}, l = 3, \cdots, p'.$$

Denote by A the $q \times p'$ matrix formed by the column vectors $(a_{I_1}, \cdots, a_{I_{p'}})$ in the ordered sequence $I = (I_1, \cdots, I_{p'})$, and $B = (b_1, \cdots, b_{p'})$. Let $W_{p' \times p'}$ represent the transformation matrix between A and B, where B is orthonormal. Therefore, the matching pursuit model is formulated as

$$Y = B X^t = \left(A \; W\right) X^t = AW$$

which provides a solution to $||Y - AX^t||$, where $W = W X^t$ represents the coefficient vector expressed in terms of the basis A.

The formulation of this model is grounded in AFD theory and derives its strength from effectively modeling nonlinear systems. Unlike conventional methods constrained by predetermined basis functions, our matching pursuit model dynamically selects the most appropriate functions from a dictionary to efficiently capture the complexities of nonlinear signals. Furthermore, as nonlinear systems typically contain substantial redundant information, our approach employs the Matching Pursuit algorithm to iteratively approximate the signal. By selectively incorporating only dictionary elements that significantly contribute to the signal's energy, we achieve a sparse approximation. Additionally, by leveraging high-order polynomials, we incorporate synergistic interactions between features through the inclusion of feature products (higher-order features). With this extended dictionary comprising higher-order features, the antigenic distance can be precisely determined and subsequently used for prediction through the sequential maximal selection of columns in the consecutively obtained orthogonal complements.

Prediction Procedure

Suppose we have two pairs of data, consisting of sequence data and antigenic data, denoted as (A, Y) and (A_1, Y_1), respectively. We use (A, Y) as the training set and (A_1, Y_1) as the testing set. Based on the matching pursuit model [3] established in the previous section, once the training set has been processed, we can derive two key outputs: the parameter set $W = W X^t$ and the index set $I = (I_1, \cdots, I_{p'})$.